THE INFORMATIONAL CONTENT OF PERCEPTUAL EXPERIENCE

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PHILOSOPHY
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Alistair Maurice Carter Isaac
July 2010
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Kenneth Taylor, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Patrick Suppes, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Persi Diaconis

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

John Perry

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Thomas Ryckman

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.
Abstract

This dissertation develops a naturalistic theory of perceptual content based on the insight that our perceptual experience measures the world. This theory is motivated in detail by examples from color perception. The basic idea is that possible color experiences define a measurement scale, and the causal relationship between this scale and the world determines the contents of the measurements it performs. This content is informational content in the sense that it captures the complete information carried by the percept about the world.

In order to have a full blown theory of perceptual content, we need to account for misperception, or the tokening of a percept in error. In order to account for perceptual error, we need to move beyond informational content and single out one amongst the many pieces of information a percept carries as its unique intentional content. The most popular strategy for this amongst naturalistic approaches to content is teleosemantics, which derives intentionality from representational function, which is in turn derived from the evolutionary history of the representing structure. My own project revisits the insight of Dretske (1981), namely that considerations of information flow alone may be enough to single out intentional content. The essential suggestion I make is that information theory provides the apparatus for analyzing representational efficacy. The intentional content of a percept may then be identified with the unique state of the world it represents most effectively.
Preface

Any project in the philosophy of psychology would do well to remember the presuppositions of psychology outlined so clearly by William James in the preface to his seminal work, *The Principles of Psychology*:

> Psychology, the science of finite individual minds, assumes as its data (1) thoughts and feelings, and (2) a physical world in time and space with which they coexist and which (3) they know. Of course, these data themselves are discussable; but the discussion of them (as of other elements) is called metaphysics and falls outside the province of this book. (James, 1890, vi)

The present work takes James’ delineation as a starting point, and combines it with a particular set of constraints and a particular methodology.

The constraints are those of naturalism, in a particular flavor which I like to call *minimal naturalism*. Some versions of naturalism endorse a robust metaphysical thesis: only the objects of fundamental physics really exist. This *physicalism* often goes hand in hand with reductionism: all natural phenomena are reducible to configurations of fundamental particles. Minimal naturalism does not endorse these strong metaphysical theses. Crucially, it does not treat physics as prioritized over any other scientific field. In the context of the present project, the most important constraint imposed by minimal naturalism is a methodological one: all explanations of natural phenomena must ultimately derive from the description of the world delivered by natural science.

The methodology is that of the semantic approach in the foundations of science, as outlined and defended by Patrick Suppes (1960, 1962, 2002). Suppes argues that
scientific theories can be usefully treated as models in the set-theoretical sense. (This contrasts with later advocates of the semantic approach, who often weaken the definition of model, allowing non-formal objects to also count as models of theories.) Furthermore, foundational questions in philosophy of science can be analyzed as questions about how a particular model should be axiomatically defined or how two distinct models (e.g. those for Newtonian mechanics and general relativity) relate. This methodology allows for philosophical questions in the foundations of science to be made mathematically precise and analyzed with formal tools.

Putting James’ insight together with the constraints of naturalism and the methodology of the semantic approach, I can now characterize the present project. I am interested in perceptual content, so this means investigating the relationship between a certain subset of James’ “thoughts and feelings” (in particular that which constitutes perceptual experience) and a certain subset of objects or properties in the physical world (in particular, those which are perceived). In both cases, we model this subset with a set-theoretical structure derived from the best current theory of the relevant science. In the case of perceptual experience, the relevant science is psychophysics; in the case of the physical world, the relevant science may be physics, chemistry, optics, materials science, or some other field depending upon the exact nature of the analysis.

The conclusion of the present work, namely that perceptual experience measures the world, is a consequence of the application of this method. Once we have a model of perceptual experience (e.g. of color) and a model of the relevant properties in the physical world (e.g. surface reflectance properties), then we can see that the formal relationship which holds between them is one of homomorphism, and in particular a homomorphism which satisfies all the properties of a measurement.

Although this particular combination of methodology and constraints is not independently defended in the present work, I hope its success in providing insight into the informational content of perceptual experience sufficiently justifies its application.
This dissertation would not have been possible without the support of my family, my friends, and the Stanford philosophy community. First and foremost, I would like to express the profound debt of gratitude I owe to the members of my committee, each of whom has made an indelible mark upon my intellectual character.

My principal co-advisors, Pat Suppes and John Perry, have demonstrated Herculean patience and faith in their dealings with me. Pat has helped me to understand the standards of true scholarship, he has shaped my understanding of science, and he has mentored me personally and professionally in more ways than I can possibly list here. He has set an impossible example (which I nevertheless aspire to emulate) with the quality of his work, the ethos with which he approaches it, and his constant dedication to pedagogy. John has had the patience to deal with my impatience, to guide me and keep me on track, even when I needed instruction in the most basic of scholarly skills. In many respects, the goal of this dissertation is to convince John Perry and, in my mind at least, this is the appropriate measure of the extent to which it succeeds or fails. Without the two of them, I would not be the philosopher I am today.

Ken Taylor and Tom Ryckman have also had an immeasurable influence on my intellectual development. Ken’s vision of a grand naturalistic synthesis in the study of mind has been an inspiring foil against which to assess my own efforts. Tom has guided me through the deepest waters of philosophy of science, always emphasizing the importance of a true scientific understanding in order to do good philosophy. Finally, Persi Diaconis’ influence is perhaps the least obvious in the text of this dissertation itself. However, his influence on my thinking about the world, and in
particular the nature of probability, has been profound. Classes and conversations with him have opened up a whole new world of mystery for me, one which I have barely even begun to explore.

In addition to my committee, many other members of the Stanford faculty have contributed to my development as a scholar and shaped this project. My understanding of philosophy and the nature of academia has been deeply enriched by the kind words, helpful comments, and exceptional instruction of Dagfinn Follesdal, Sol Fefferman, Lanier Anderson, Nadeem Hussain, Brian Skyrms, Chris Bobonich, Michael Friedman, Tom Wasow, and David Israel. I have been fortunate enough to have many fruitful interactions outside the philosophy department as well, especially those with Michael Ramscar, Ivan Sag, Stanley Peters, Lauri Karttunen, and Pat Langley, whose constructive outlook and broad knowledge base have given me hope in the possibility of cross-disciplinary communication.

David Hills deserves special mention for many helpful exchanges, and a particularly interesting conversation about the nature of color which we managed to maintain over the course of many separate bus rides together. George Smith, while only a visitor at Stanford, nevertheless shaped my understanding of the nature of science in a profound way. Finally, I am deeply grateful and appreciative for all that Johan van Benthem has kindly contributed to my training and development. He has taught me, mentored me, guided me professionally, and even supported me financially on two research trips to Europe, both of which had a substantive effect on the contents of this dissertation. Not only his ideas, but also the opportunities he created for me have deeply affected this project.

I would like to thank my peers in the philosophy community as well, especially the members of my cohort: Alexei Angelides (who taught me the importance of history), Jesse Alma, Quayshawn Spencer, Dan Giberman, Tomohiro Hoshi, Peyton McElroy, and Johanna Wolff. Johanna deserves special thanks for her emotional and moral support, as well as many long, illuminating conversations. Teru Miyake and Donovan Wishon have kindly read and commented upon previous drafts of this project; the substantial overlap in interests between mine and theirs has produced a number of deep and constructive interactions. Additionally, conversations and reading groups
with Angela Potochnik, Thomas Icard, Tal Glezer, Assaf Sharon, Ben Wolfson, Darko Sarenac, Patrick Girard, Tyler Greene, Audrey Yap, Joel Velasco, Darren Bradley, and Eric Pacuit have helped to shape my views. And finally, amongst all these stimulating friends, I must single out Tomohiro Hoshi whose unfailing support and frequent collaborations (on everything from golf to game theory) have enriched my Stanford experience beyond measure.

The last two years of my life have been dominated by the class *HSC: Great Ideas, Observations, and Experiments in the History of Science*, which I designed and taught for the EPGY Online High School. Pat Suppes encouraged me to take the job and guided me in designing the course, and Tomohiro Hoshi was present every step of the way, first assisting with the research and ultimately taking over teaching duties. Many of the tidbits of scientific history to be found in the following pages were discovered while doing research for HSC. More importantly, however, I had the privilege of teaching a number of amazing students these past two years: their enthusiasm for science and philosophy was a constant inspiration and reminder of the importance of pedagogy.

This list only begins to scratch the surface of the rich experience I have had here at Stanford; I apologize to anyone whom I may have inadvertently forgotten. Saving the most important for last, I would like to thank my Father for his unwavering love and support during the difficulties of graduate school and the dissertation writing process.
## Contents

Abstract iv

Preface v

Acknowledgements vii

1 From Information to Intentionality 1
  1.1 Introduction 1
  1.2 From Continuous to Discrete Representational Systems 4
  1.3 From Information to Intentional Content 7
  1.4 Note on Propositions and Content 18
  1.5 Plan of the Work 21

2 The Hierarchical Structure of Perception 24
  2.1 Introduction 24
  2.2 Psychophysics and Beyond 26
    2.2.1 The Science of Experience: An Example from Weber 26
    2.2.2 Hierarchical Consciousness in Philosophy and Neuroscience 30
  2.3 What is Color? 34
  2.4 Transduction and Scotopic Vision 41
    2.4.1 Transduction and Information 41
    2.4.2 Transduction in Rod Cells 43
  2.5 Pixel Color Perception in Photopic Conditions 47
    2.5.1 Transduction in Cone Cells 48

x
3 Perception as Natural Measurement

3.1 Introduction ........................................... 65
3.2 Measurements and Mappings .......................... 67
  3.2.1 What is Measurement? ............................ 67
  3.2.2 Content as Invariance under Suitable Transformations .... 71
  3.2.3 Representational versus Artifactual Structure ............ 75
3.3 Distortion in Natural Measurement .................... 79
  3.3.1 Natural Measurement .............................. 79
  3.3.2 The Objection from van Fraassen .................. 81
  3.3.3 Turning van Fraassen on his Head: Distortions as Evidence ... 86
3.4 The Problem of Error, part 1 .......................... 91
  3.4.1 Precision and Accuracy ........................... 91
  3.4.2 Inaccuracy in Perception ......................... 94
3.5 Conclusion ......................................... 98

4 Discretization and Detachability ......................... 99

4.1 Introduction .......................................... 99
4.2 The Problem of Error, part 2 .......................... 101
4.3 Gärdenfors' Voronoi Tessellations .................... 108
4.4 Recent Empirical Evidence for the Gärdenfors Model ........ 117
4.5 Conclusion ......................................... 128

5 The Informational Content of Natural Signs ............. 130

5.1 Introduction .......................................... 130
5.2 The Problem of Error, part 3 .......................... 132
5.3 Basic Concepts of Information Theory ................ 140
  5.3.1 From Probability to Information .................. 141
# List of Figures

1.1 The lay of the land. .................................................. 8  
1.2 The plan of the work ................................................ 22  

2.1 Sequence of structural changes as rhodopsin absorbs light. ........ 44  
2.2 Normalized Rod Sensitivity. ......................................... 45  
2.3 Cone response to different wavelengths of light. ................. 50  
2.4 Sensitivity functions for S, M, and L cones. ....................... 51  
2.5 Setup for a color matching experiment. ............................ 53  
2.6 Physiological and behavioral responses to color compared. ........ 56  
2.7 CIE 1931 color space. ............................................... 57  
2.8 Data from Wright (1941) plotted in the CIE 1931 color space. .. 60  
2.9 Color space of Kirschmann (1895) ................................ 61  

3.1 Two Types of Mapping Relation .................................... 75  
3.2 The Müller-Lyer Illusion ............................................. 96  
3.3 Context as calibration on color scales. ............................ 97  

4.1 The functional architecture of the mind. .......................... 102  
4.2 The functional relationship between concepts and properties. ..... 107  
4.3 Network activation as vectors in a state space. .................... 109  
4.4 Vowels graphed into formant space. .............................. 112  
4.5 Tessellation of the space of stop constants. ....................... 113  
4.6 The subspace of skin colors. ....................................... 115  
4.7 The cylindrical color space implied by the WCS stimuli. ........ 118
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.8</td>
<td>$\beta$-mappings amongst systems of color terms.</td>
<td>120</td>
</tr>
<tr>
<td>4.9</td>
<td>Munsell stimuli from WCS mapped into CIELAB.</td>
<td>123</td>
</tr>
<tr>
<td>4.10</td>
<td>Well-formedness of Berinmo color categories.</td>
<td>127</td>
</tr>
<tr>
<td>5.1</td>
<td>Two probabilistic processes.</td>
<td>147</td>
</tr>
<tr>
<td>6.1</td>
<td>Twelve metameric surface reflectance profiles.</td>
<td>182</td>
</tr>
<tr>
<td>6.2</td>
<td>Churchland’s canonical approximation ellipses.</td>
<td>186</td>
</tr>
<tr>
<td>6.3</td>
<td>Frequency and location of metameric curve intersections.</td>
<td>189</td>
</tr>
<tr>
<td>6.4</td>
<td>Rough schematic of the causal path resulting in color experience.</td>
<td>195</td>
</tr>
<tr>
<td>6.5</td>
<td>Preservation of distal information in causal path.</td>
<td>197</td>
</tr>
<tr>
<td>6.6</td>
<td>Measuring the representational efficiency of color perception.</td>
<td>200</td>
</tr>
</tbody>
</table>
Chapter 1

From Information to Intentionality

1.1 Introduction

What does our perceptual experience tell us about the world? This project will develop a general apparatus for answering this question and apply it in detail to the specific example of color perception. The theory developed here is an instance of a naturalistic approach to content, i.e. I will derive an analysis of the content of perceptual experience from descriptive facts about the world as delivered by natural science.

The greatest challenge for a naturalistic approach to content is deriving intentionality. Intentionality is a complex concept. Here, I will restrict attention to but two of its many facets: intentional content is unique and an intentional representation may be tokened in error, i.e. it may misrepresent. By “unique” I just mean that intentional content is about one thing in particular. In general, intentionality may be cashed out as aboutness or directedness, but all more general approaches to analyzing intentionality imply these two properties—that intentional content is unique and that it may be in error.

Naturalistic theories of content have found it relatively easy to define (what I will call) informational content. Usually, we think of the content of a representational state as a proposition (however, see discussion of the role of propositions in Section 1.4). But an event in nature may contain information about many different events in its
causal past. Correspondingly, if we derive content from the information contained in an event, it looks like we will get a set of propositions, rather than a single one. This set of propositions is the informational content of the event. How can we single out one of these propositions as the unique intentional content of this event?

Furthermore, if our account of content is derived from the causal history of an event, it appears as if the proposition we ascribe as its content can never be in error. Yet this is one of the critical desiderata for providing an account of intentional content.

In response to these two problems, naturalists have turned to a teleological approach. In general, this approach has two components. First, mental states represent the proposition which it is their function to represent. This analysis succeeds in singling out one proposition as the content, and it allows for error since content no longer depends upon a direct causal relationship with the world. However, appeal to function simpliciter does not satisfy the constraints of naturalism. In order to connect representational function to the natural world, the second component of the teleological approach identifies function with the representational role a mental state evolved to serve.

I believe the teleological approach is fundamentally misguided. This, then, provides one major motivation for the project:

Q1 How can a naturalist derive intentional content from informational content without appealing to teleology?

The background behind this problem will be discussed in more detail in Section 1.3. In a nutshell, my solution will depend upon two key ideas. First, there are actually several different types of error, some of which can be accommodated naturally within the information theoretic framework. Second, the key source of the error commonly referred to by terms such as “misrepresentation” depends upon the detachability of the semantics for low-level perceptual states from that of higher-level perceptual states. In my account, the content of low-level perceptual states is determined by causal interactions with the world, but the content of higher-level states is determined by their relationship to lower-level states. This split level account will allow for the detachability of, say, color terms from color experience.
This brings us to the second problem which motives this project, which is derived in part from the literature on nonconceptual content. One of the key data points in that literature is the supposed richness, or fineness of grain, of perceptual experience. This richness is contrasted with the discrete, easy categorization found within conceptual content. I do not intend to tackle the question of what is properly understood as conceptual and what as nonconceptual. However, I do see a simpler and clearer problem underlying the debate, namely the relationship between continuous and discrete representational systems.

A representational system is a set of mutually exclusive elements which share content about a single domain. For example, words in a language constitute a representational system. One reason we might think perceptual experience has a special type of content is that it appears to be able to vary continuously, unlike words, categories, and concepts, which are in general discrete. I will argue that we can find both discrete and continuous representational systems within perceptual experience. This inspires the following question:

**Q2** How do the continuous and discrete aspects of experience relate? How do their contents relate?

The background behind this issue will be discussed in more detail in Section 1.2.

The claim that perceptual experience can involve both continuous and discrete representational systems may seem puzzling. Before we can investigate **Q1** and **Q2**, then, we must first be clear on

**Q3** What is perceptual experience?

I contend that perceptual experience is too complex to be treated as a single object. Rather, there are many facets of perceptual experience and we can study each of these facets only to the extent that we can isolate them in an experimental setting. In Chapter 2, I will argue that perceptual experience is a rich, hierarchical structure, each level of which can be isolated in conscious experience through suitable experimental design, and corresponds to some level in the hierarchy of neural processing of sensory input.
Throughout this project, I will use the perceptual experience of color as a guiding example. Color experience is particularly suited to investigating the questions outlined above. For example, it is relatively easy to isolate and study early levels of the hierarchy of color experience. There is a broad literature on the relationship between the continuous space of possible color percepts and the discrete space of color terms. Finally, the determination of the intentional content of color experience (in the sense I use the term here) has been much debated in the literature on metamers.

The following two sections provide background to Q1 and Q2. Then I will briefly discuss the interpretation of propositions relevant to the present project. In the final section of this chapter I will give the plan of the work and explain the organization of the remaining chapters.

1.2 From Continuous to Discrete Representational Systems

The idea that there may be a type of content distinct from that we find in utterances and beliefs arises from the consideration of the difference between what we can experience perceptually and what we can remember and communicate. Although the first use of the term “nonconceptual” to characterize this distinctive type of content is probably by Gareth Evans (1982), we can already find the essential points in Dretske’s 1981 discussion of sensory vs. cognitive processes:

Sensation, what the ordinary man refers to as the look (sound, smell, etc.) of things, and what the psychologist refers to as the percept ... is informationally profuse and specific in the way a picture is. Knowledge and belief, on the other hand, are selective and exclusive in the way a statement is. ... Our sensory experience embodies information about a variety of details that, if carried over in toto to the cognitive centers, would require gigantically large storage and retrieval capabilities. There is more information in the sensory store than can be extracted, a limit on how much of this information can be exploited by the cognitive mechanisms. (Dretske, 1981, 142)
The general features of perception identified here by Dretske have motivated much of the literature on nonconceptual content. As summarized by York Gunther in the introduction to the seminal anthology on the topic:

Many proponents of nonconceptuality are deeply impressed by the specificity of its content, its richness of detail, and fineness of grain. . . . Through vision, for example, we are capable of discriminating shades of color and shapes of objects that seem to outrun the descriptions and categories available to us. Or when we hear a sound coming from a certain direction, we tend to turn our heads to look for its source without thinking or calculating, inferring or deducing where to turn. . . . Like our perception of color and shape, our experience of space is believed not only to be immediate but indeterminate and even inexpressible. (Gunther, 2003, 3)

However, despite the intuitive strength of these motivating examples, the debate about nonconceptual content is plagued by disagreements about how exactly the term “conceptual”, and consequently “nonconceptual”, should be understood. Many participants in the debate readily acknowledge this point and begin their contributions by rejecting the definitions of others or stipulating their own. For example, lecture 3 of McDowell (1994) rejects all Evans’ attempts to distinguish nonconceptual content, essentially arguing that content is conceptual “all the way down”. In contrast, Stalnaker (1998) begins with his own definitions of conceptual and nonconceptual, ultimately concluding that content is nonconceptual “all the way up” (106).

In the present work, I would like to set aside the issue of how to distinguish the distinctively conceptual from the nonconceptual. However, I am interested in how to provide an account of content for perceptual experience which is sensitive to its richness of detail and fineness of grain. Furthermore, I take it this account will differ substantially from the correct account of content for discrete spaces. So, in the present project, the content of the experience of color and the content of color terms will be treated differently. Essentially, color experience derives its content from its causal relationship to the world. This relationship will be treated as a measurement, and we will see in Chapter 3 that the theory of measurement provides an analysis of representation which applies directly to our experience of color (and mutatis mutandis to other spaces of perceptual experience).
Characterizing perceptual experience of color in terms of a continuous space and claiming that the relationship between this space and the world is representational may appear to be an implicit endorsement of nonconceptual content. However, the view presented here should stand or fall on its success as an account of perceptual experience and what it tells us about the world, not any analysis of what we mean (or don’t) by “conceptual”. Furthermore, it may appear that this is an abuse of the term “representation”; in particular, many philosophers have insisted that “there is no representation without misrepresentation”, yet if the representational relationship between mind and world is simply causal, there seems to be no room for error and, correspondingly, no possibility of misrepresentation.

In the present work, I interpret representation in a very thin sense: to say that $X$ represents the proposition $p$ is just to say that the information that $p$ is presented again in $X$ (modulo the qualifying discussion of propositions in Section 1.4); to say that $x$-ness represents the property of $q$-ness is just to say the features of $q$-ness are presented again by $x$-ness. This locution may seem funny when speaking of properties, but in the case of both measurement and perception, the features of a property are cashed out in terms of the position it serves within a structured set of properties of which it is a member. Lengths, for example, form such a set, and the structure of this set is the natural ordering over lengths from lesser to greater. So, we can represent the property of having a specific length by any property we want, so long as that property is also part of a well defined family of properties which is structured by an ordering with the same features as the ordering over lengths. A representation theorem shows that a particular set may be used for this purpose. In the case of lengths, we’d like to prove that the positive real numbers may be used as the representing set. In the typical case, there may be many ways to assign numbers to a measured quantity, and the interesting work comes in investigating the invariances across all such representations for a particular measurement procedure.

Notice, however, that only the thin notion of representation is needed here. Of course, in the case of humans performing measurements, there is always the possibility that they apply the measurement procedure incorrectly and thereby misrepresent the intended property. This possibility for error arises due to how the two domains
(say, lengths and numbers) are connected in practice. But it is not in virtue of this practice, but rather in virtue of the formal theorem, that numbers constitute a legitimate representation of lengths. In fact, notice that one cannot make an error in measurement without an antecedent thin representational relationship between the object domain and the representing domain. If I perform a procedure I claim is a measurement and say it returns the value 4, but the procedure is not precisely defined or coherent enough that a representation theorem could (at least in principle) be proven, then there is no sense in which my “measurement” can be in error. I appear to have simply stipulated that 4 represents the intended quantity without actually performing a measurement in the strict sense.

So, the possibility of error is not constitutive of representation in this thin sense. Nevertheless, I will argue in Chapter 3 that even when representation is cashed out in terms of causal relationships, error can emerge. The trick here is that there are actually several different types of error. Error in the sense of imprecision, for example, is always present in natural measurements such as that made by the perceptual system. Error in the sense of inaccuracy is harder, but not impossible, to derive naturalistically. The real trick is to provide an analysis of content such that representations can be detached from the causal chain from world to experience and be tokened in error. A thick notion of representation demands that such mistokenings be possible, and a convincing analysis of intentional content must provide a semantics that allows for detachability. This problem is addressed in Chapter 4.

1.3 From Information to Intentional Content

Before discussing how the problem of intentional content emerges for the naturalist, it will be helpful to get a feel for the general lay of the land. Amongst naturalistic accounts of content, there are two major strategies: teleological and information theoretic. I take causal approaches to fall under the information theoretic account, for reasons to be discussed below. The general idea behind the information theoretic approach is to derive the content of a representational state from the information it contains. The general idea behind the teleological approach is to derive the content of
a representational state from the representational function is has evolved to perform. This approach is often called “teleosemantics”.

A color “realist” claims that colors are properties of surfaces. Strictly speaking, the “realism” debate in the philosophy of color literature cuts across the information theoretic / teleological distinction. For example, Byrne and Hilbert, two of the most prominent color realists of recent decades, are explicitly agnostic about teleological versus causal approaches to content. However, this agnosticism is not sustainable for a general theory of perceptual content.

One reason color realists can ignore general questions concerning naturalistic content is that their debate is focused exclusively on the relationship between color experience and surface properties. They assume as the starting point of this debate that color experience is about surfaces. Consequently, the main challenge for an information theoretic approach (which a teleological approach attempts to solve) simply does not arise.

---

1The color realism debate is discussed more thoroughly in Chapter 6.
CHAPTER 1. FROM INFORMATION TO INTENTIONALITY

This challenge is the problem of deriving intentional content from informational content. This challenge has two components. The first component is singling out \textit{one} amongst the many pieces of information which can be extracted from an event as its unique content, call this the problem of \textit{uniqueness}. The second component requires that we explain how this event can be tokened in error, i.e. without direct causal stimulation from that situation which we take it to uniquely represent. Call this the problem of \textit{detachability}. Let's examine how these problems arise within the information theoretic account.

As naturalists, we begin by helping ourselves to the causal story told us by science. For any perceptual experience, we can trace the chain of causality leading from external stimuli to a conscious perceptual experience. Of course, there are details of this chain which we do not understand, in particular, the relationship between neural firing and conscious experience. However, the assumption that there is a causal relationship here is a starting point for the scientific investigation of experience, and thus must be assumed if we are to use the science of color vision as the starting point of our analysis.

Now, it is natural to assume that color experience tells us something about objects or events in this causal chain. However, simply knowing the causal chain will not be good enough to fix content. For one thing, the entire past light cone of an event is potentially part of its causal past. However, we cannot learn about the entire past light cone of an event just from witnessing it. This is why any causal theory of content requires an analysis of information. The only candidates for the contents of a perceptual experience are propositions concerning those events or objects which fall in its causal past \textit{and about which it contains information}.

So far so good. And, as we will see in Chapter 5, information theory provides enough apparatus to define the informational content of a perceptual experience by combining the total information it contains about all events potentially in its causal past into a single formal object. However, in general, this will not be enough to single out a unique event in the past of a perceptual experience as its intentional content. For example, consider the case of color perception. Light is radiated from a source; it causally interacts with the physical structure of a surface; the changed light
is absorbed at the retina. Now, my experience of color contains information about all the steps in this process; I can learn about the nature of the illuminant, surface properties, neural activity at my retina, etc. from an experience of color.

The most substantial attempt to develop the information theoretic approach to content without full-blown teleology is Dretske (1981). Dretske’s analysis runs into the exact problem discussed here: \textit{information is profligate, but intentionality is demure.} (Here he uses “meaning” for what I have been calling “intentional content”.)

The distinction between meaning and information becomes even more evident when we examine cases of nomically nested information. Assuming it to be a law of nature that water expands upon freezing, no signal can carry the information that some body of water is freezing without carrying the information that this body of water is expanding. But the statement, “This body of water is freezing” can \textit{mean} that this body of water is freezing without \textit{meaning} that this body of water is expanding.

It may seem as though the informational content of a signal is threatening to overflow. Our discussion has revealed that when a signal $r$ carries the information that $s$ is $F$, it also carries the information that $s$ (or $t$) is $G$ when the fact that $s$ (or $t$) is $G$ is nested (either analytically or nomically) in $s$’s being $F$. Signals, it seems, are pregnant with information. And so they are. (Dretske, 1981, 72–3)

(In Dretske’s terminology, the fact that $s$ is $G$ is \textit{nested} in a signal carrying the information that $s$ is $F$ if the fact that $s$ is $F$ itself carries the information that $s$ is $G$ (for example, if being $G$ is a nomic consequence of being $F$).)

Dretske solves the problem of uniqueness in three steps. In his definition of information, he stipulates that a signal can only carry the information that $s$ is $F$ if the conditional probability of $s$ being $F$ given the signal is tokened equals 1. On Dretske’s analysis of probability, this can only obtain if there is a nomic dependence between $s$ being $F$ and a tokening of the relevant signal. Suppose $G$ is extensionally equivalent to $F$, yet not for nomic reasons, but as a mere accident of correlation. For example, suppose I see the words “Byron was here” carved into the side of a Greek temple. Now, Byron visited many temples in Greece and carved his name into many of them as well. It may be that the temples Byron has visited and the temples upon which “Byron was here” have been carved are extensionally equivalent. However, if
this carving of “Byron was here” was not causally connected to Byron’s presence at the temple (say, it was carved many years later by a Byron admirer, uncertain of his idol’s path through Greece), then this extensional equivalence, although perfect, is accidental, and thus, on Dretske’s definition, the signal “Byron was here” does not carry the information that Byron was at the temple (Dretske, 1981, 72–7).

Ruling out extensionally equivalent properties is not enough to get uniqueness, however. Dretske also must deal with the fact that an event can contain information about multiple stages of its causal past. He considers as an example, the ringing of a doorbell and our perceptual experience of it. He argues that the ringing of a doorbell contains the information that there is a person at the door. However, the content of the perceptual experience is not that there is a person at the door, but that the doorbell is ringing. Dretske explains this by appealing to a notion he calls primary representation.

\[ S \text{ gives primary representation to property } B \text{ (relative to property } G) = S' \text{’s representations of something’s being } G \text{ depends on the informational relationship between } B \text{ and } G \text{ but not vice versa.} \]

Our auditory experience represents the bell ringing and it represents the button’s being depressed. But only the former is given a primary representation because the information the experience carries about the depression of the button depends on the informational link between the button and the bell while its representation of the bell’s ringing does not depend on this relationship. If we short-circuit the doorbell wires . . . the information tie between the bell and the button is broken. When this tie is severed, the auditory experience continues to represent (carry information about) the ringing bell, but it no longer carries information about the depression of the button. (Dretske, 1981, 160)

Third, Dretske addresses the problem that a precise piece of information also carries corresponding coarser pieces of information. A photo of a cup with coffee in it carries the information that there is coffee in the cup, but it also carries arbitrarily precise information about the exact shape of the cup, the exact quantity of coffee, etc. Dretske addresses this issue with his distinction between analog and digital (note: his usage of these terms is highly idiosyncratic and does not conform to standard usage).
This distinction allows Dretske to deal with the problem noted in the previous section of how the fine-grained features of perceptual experience are related to the more coarse categories we find in beliefs and language.

I will say that a signal (structure, event, state) carries the information that \( s \) is \( F \) in digital form if and only if the signal carries no additional information about \( s \), no information that is not already nested in \( s \)’s being \( F \). If the signal does carry additional information about \( s \), information that is not nested in \( s \)’s being \( F \), then I shall say that the signal carries this information in analog form. When a signal carries the information that \( s \) is \( F \) in analog form, the signal always carries more specific, more determinate, information about \( s \) than that it is \( F \). (137)

Dretske posits an analog to digital converter in the information processing chain within our brains. It is at this barrier that rich and fine-grained perceptual experience is translated into the coarser categories found in beliefs and communication. This is the barrier between the sensory and the cognitive for Dretske (between the nonconceptual and the conceptual in the terminology of the previous section).

Finally, how does Dretske explain detachability? Here, his story depends upon learning. We learn a symbol, say, the word “red”, in a training environment. Once we have the symbol and have left the environment, however, we can apply it in inappropriate circumstances.

\[ \ldots \text{once the subject has articulated a structure that is selectively sensitive to information about the } F\text{-ness of things, instances of this structure, tokens of this type, can be triggered by signals that lack the appropriate piece of information. When this occurs, the subject believes that } s \text{ if } F \text{ but, because this token of the structure type was not produced by the information that } s \text{ is } F, \text{ the subject does not know that } s \text{ is } F. \text{ And if, in fact, } s \text{ is not } F, \text{ the subject falsely believes that } s \text{ is } F. \text{ We have a case of misrepresentation—a token of a structure with a false content. We have, in a word, meaning without truth.} \]

(195)

How successful is Dretske’s account? First, the analysis of detachability is underdeveloped; this is an issue Dretske deals with more thoroughly in later works, though these all fall within the framework of teleosemantics. We will attempt to do better,
without the teleology, in the present work. As for Dretske’s account of uniqueness, the
analysis provided in Chapter 6 will draw inspiration from Dretske’s general strategy
for singling out one stage in the causal chain (his “primary representation”), though
it will differ in the details due to our different treatments of probability.

One presupposition of the present analysis is that perceptual properties are exten-
sionally defined. This means that the problem of ruling out extensionally equivalent
properties in determining intentional content simply does not arise. The worry that
we need to distinguish extensionally equivalent properties arises naturally if we con-
sider natural language, but it simply does not make sense when investigating the
informational relationship between perceptual experience and the world. On my
view, the order of explanation is reversed. We first need to explain the detachabil-
ity of representations from our causal connection to the world before we can make
sense of the idea that there might be extensionally equivalent, yet distinct, properties
represented. To say it another way, there are no instances of extensional equiva-
ience amongst perceptual properties, i.e. those properties directly perceived by our
perceptual apparatus.

Finally, the problem of analog to digital conversion (in the traditional sense of the
terms) will be dealt with in Chapter 4. However, the worry there is not Dretske’s,
which rests on a misunderstanding about how analog systems represent. Dretske
seems to assume that an analog representation can represent to an arbitrary degree of
precision (hence the stipulation that analog signals “always carr[y] more specific, more
determinate, information”). However, as we will see in Chapter 3, an analog system
(taken in the traditional sense of a continuous system) has limits in its representational
power, which we measure in terms of the precision of the measurement it performs.
This precision (or lack thereof) can be directly inherited by the corresponding discrete
representation after analog to digital conversion.

The full-blown teleological approach (the most famous example, perhaps, being
Millikan, 1984) appeals first to evolutionarily evolved function and from this de-
rives both the uniqueness and the detachability of intentional structures. After 1981,
Dretske switches to a full-blown teleological approach. In 1988, for example, he de-
fines content in terms of representational function:
If an RS [representational system] has the function of indicating that \( s \) if \( F \), then I shall refer to the proposition expressed by the sentence "\( s \) is \( F \)" as the content of the representation. (Dretske, 1988, 70)

Furthermore, functions in natural systems are analyzed by analogy with the functions of artifacts, and the role played by design in the case of artifacts is played by evolution in the case of natural systems. Dretske (1988) emphasizes that only with an analysis of representational function can one provide an analysis of representational error.

Misrepresentation depends on two things: the condition of the world being represented and the way that world is represented. The latter, as we have seen, is determined, not by what a system indicates about the world, but by what it has the function of indicating about the world. And as long as there remains this indeterminacy of function, there is no clear sense in which misrepresentation occurs. Without a determinate function, one can, as it were, always exonerate an RS of error, and thus eliminate the occurrence of misrepresentation, by changing what it is supposed to be indicating, by changing what it is its function to indicate. (69)

It is not clear to me exactly why Dretske felt the need to move to a full-blown teleological view. My suspicion is that he realized his account of detachability in Dretske (1981) was underdeveloped and inadequate to the explanatory task. Certainly, he does not present the later work as a change from his earlier position, but merely as a supplementation and elaboration. However, neither function, nor evolutionary history, play an important role in the theory developed in Dretske (1981). Perhaps he was motivated by the wide-ranging application of the teleological view developed in Millikan (1984). One appeal of the teleological approach developed by Millikan is that it addresses not only meaning, but biological categories in general. If we are worried about the function of the heart or the liver (for example, in order to identify cases where a heart or liver is not functioning correctly), then we can apply the teleological apparatus she develops. Representational function emerges then as a special case of a more general theory.

There are many reasons for skepticism about the need for or success of the teleose-mantic approach. Here are three which motivate the present project.
CHAPTER 1. FROM INFORMATION TO INTENTIONALITY

The first is skepticism that evolutionary history will determine biological function in a truly non-circular, non-vacuous manner. Certainly, Millikan (1984) develops an enormously complex apparatus to this end, and a vast number of works have followed in her footsteps. I do not pretend to refute or diminish the value of this literature. However, there are well known worries about the role of teleological explanations in evolutionary theory generally. Evolutionary biologists themselves have wrestled with the apparent circularity in some of their analyses. The seminal paper “The Spandrels of San Marco and the Panglossian Paradigm” (Gould and Lewontin, 1979), for example, challenged the adaptationist program, which seeks to explain traits in terms of their adaptation to serve a particular function. Yet adaptationism appears to be a presupposition of the teleosemantic approach. The essence of teleosemantics is to derive the intentionality of representational states from selective pressures:

In sum, if we look at the whole human person in the light of our history of evolution by natural selection, minding the continuities between humans and other animals, it appears that all levels of purpose have their origin in adaptation by some form of selection. In this sense, all purposes are “natural purposes.” (Millikan, 2004, 13)

The entire project collapses, however, if the attribution of directedness to evolutionary pressures fails. This is certainly still an open question in evolutionary theory, and the adaptationist program has recently received a rousing defense by Andrews et al. (2002). Personally, however, I am sympathetic to the commentary on this defence provided by Figueredo and Berry, who express skepticism that a unique function can ever be specified for a biological trait within the evolutionary framework:

Another troubling point has to do with the seeming anthropomorphism that has crept into the debate about the “real” function of an adaptation. We speak of Nature, personified, “designing” adaptations for a specific function, as if with a conscious intention. Nevertheless, Andrews et al. correctly point out that what distinguishes an adaptive function is that its effect “either enhances or inhibits the replicative success of the genes from which it develops” . . . This means that any effect constitutes an adaptive function if it contributes to survival and reproduction. No intelligent design is implied in natural selection. Nature is “blind” to what might have
been the “original” function of an adaptation. Beneficial genetic mutations arise by blind idiot luck and not providence. They are “selected” by Nature to increase or decrease in the population based on whatever fortuitous effects they might happen to have by purest accident. This is not only true of exaptations,\(^2\) but also of adaptation. (Figueredo and Berry, 2002, 517)

If our attribution of a kind of natural directedness to selective pressures is in error, then our attempt to derive the intentionality of content from this directedness will fail.

These worries have reemerged in the philosophy literature due to Fodor and Piattelli-Palmarini (2010), who argue that there can be no laws of natural selection, only ex post facto explanations. Fodor and Piattelli-Palmarini (FP)’s too strong claims have been swiftly rebuffed by both Block and Kitcher (2010) and Sober (forthcoming). It is important to point out that these responses do not claim as much for natural selection as is needed for the teleosemantic project to work. For example Sober criticizes FP’s discussion of the classic example of hearts and pumping blood. All hearts both pump blood and make a thump-thump sound; these traits are coextensive. FP claim that, in principle, evolutionary biology can provide no laws which explain which of these two coextensive traits is selected for. Sober responds that given two traits such as this, evolutionary biologists can design experiments to determine which is being selected for. However, this is not enough to get a unique function for every trait (the point of Figueredo and Berry, 2002), but that is precisely what is needed for teleosemantics to derive intentionality from adaptation.

The second reason for doubting teleosemantics is a problem for any historical approach to meaning or function, most famously associated with Davidson (1987). Davidson first presents the “swampman” problem, in which lightening strikes a swamp causing an exact replica of Davidson himself to emerge full formed. As Davidson points out, although the swampman is exactly like him in every detail, nevertheless, the thoughts the swampman experiences do not have meaning, and correspondingly his organs do not have functions on a teleological account. If our representational

\(^2\)“An exaptation is a preexisting trait that acquires a new beneficial effect without modification to the phenotype by selection” (Andrews et al. (2002), 490).
states represent only in virtue of their histories, then if our history were changed or removed, our representations would be also.

Now, it’s true that the swampman example is contrived and physically implausible. The intuition, however, that my brain successfully represents the world in virtue of the relationship it stands to the world right now (and not in virtue of the details of my (evolutionary) past) is a compelling one to which I am deeply sympathetic. I am simply unwilling to bite the bullet teleosemanticists must bite here. My reflection in a pond does not depend upon the history of the pond, but the physics of light. Likewise with a camera obscura—a small hole into a darkened room will project an upside down image onto the wall, no matter how or why the hole was made. Our eyes exploit the same physical principles as the camera obscura, why should their evolutionary history be the measure of their success or failure in imaging the outside world rather than their causal interactions with that world today?

This brings us to the third reason for skepticism about the teleological approach, which is really just optimism for an information theoretic approach. Dretske (1981) devotes many pages to explaining information theory, yet he hardly gets past the first chapter of Shannon and Weaver (1949), to say nothing of the extensive literature on information theory which has emerged in its wake. As has been pointed out by many commentators, Dretske abandons the probabilistic tools which lie at the heart of information theory early in his program. What if we return to information theory and see how far an analysis will get which makes use of its full apparatus? The intuition that the return to information theory may provide a fruitful alternative to teleosemantics is expressed by Godfrey-Smith (2006), although he does not pursue the suggestion in depth.

This is the motivation of the present work—to return to the roots of Dretske (1981) in information theory and see if we can produce a more satisfactory account of content by taking the mathematical theory of information seriously. If successful, this account will provide an alternative to teleosemantics for those who share my skeptical worries.
1.4 Note on Propositions and Content

At several points above, I have said that the content of a representational state is a proposition (or, for informational content, a set of propositions). This is the standard approach to content as motivated by considering natural language. In analyzing the semantics of natural language, it appears that the same content can be expressed by multiple utterances, so we define an abstract object which is invariant across them, the proposition which is expressed.

Although the appeal to propositions has, I hope, been helpful in getting our discussion off the ground, they will not play a major role in the sequel. We are interested here in the relationship between representational states of the mind and states of affairs in the world. Propositions fall in between. We could first characterize the proposition which is the content of a mental state, then give a semantics for the proposition in terms of the state of affairs it represents. However, for the present purposes it will be more useful to ignore the intermediary step and directly examine the relationship between mind and world. Of course, everything done without propositions here could easily be reconstructed in propositional talk, but my personal view is that this would obscure rather than illuminate the details of my theory of content.

A second reason for avoiding propositions is somewhat more complex and closely related to the fact that we are developing a theory of perceptual content. The essential feature of a proposition is its subject-predicate structure. Appeal to an object (e.g. “chair”) or to a property (e.g. “is red”) by itself does not make a claim about the world and, consequently, cannot serve the role of content. We must combine an object with a property in order to form the simplest type of propositional structure (e.g. “the chair is red”). This general idea is what motivates Dretske’s talk of a signal containing the information that “s is F”, where s is just some object and F is a property.

In order to take full advantage of information theory and the theory of probability upon which it rests, however, it will be more useful to think of content in terms of a random variable taking on a particular value. A random variable is just a function defined on a sample space. In order to get a probabilistic analysis off the ground,
we need to define the space of possibilities under consideration, this is the *sample space*. For example, if we roll a die, then the sample space is just the space of possible outcomes, i.e. \( S = \{1, 2, 3, 4, 5, 6\} \), and the natural probability distribution \( p \) assigns an equal probability to any of these outcomes. We can then define different random variables over this sample space depending upon the outcomes of interest to us. For example, we could define the random variable \( X \) over the range \{*even, odd*\} such that \( X(1) = X(3) = X(5) = \text{odd} \) and \( X(2) = X(4) = X(6) = \text{even} \). Then the probability of \( X \) taking on the value *even* is defined in the natural way, namely \( P(X = \text{even}) = p(2 \lor 4 \lor 6) = 1/2 \).

I claim that treating content in terms of random variables taking on particular values satisfies the essential intuition that motivates treating content in terms of propositions, but is more general than the object–property framework to which Dretske restricts himself. The essential intuition here is that we need content to be a claim about the world that can be true or false. The claim that the die comes up even (i.e. that \( X = \text{even} \)) is just such a claim. The random variable framework is more general than the object–property framework, however, because it applies to any procedure with a well defined outcome. This can be the outcome of an experiment or a measurement, for example, *even when there is debate about what objects are involved*.

This move receives additional motivation from worries that have emerged in analyzing the content of perceptual experience. The early attempts by Russell, Ayer, and others to cash out the content of experience in terms of sense data ran into serious problems. These philosophers attempted to cash out sensory experience in terms of simple propositions, e.g. “this patch is red”. One of the many problems with the sense data approach is that it seems to presuppose objects of questionable ontological status (patches in the visual field, etc.). One response to this problem comes from the adverbial theory, which argues that perceptual experience should be understood in terms of adverb-like modifications in quality (rather than attributions of properties to questionable objects). For example, rather than saying my experience has as content the proposition “this patch is red”, an adverbialist might say I am experiencing “redly”.
Personally, I find the talk of adverbs here confusing and unintuitive. However, I am sympathetic to the basic intuition that we should be able to cash out low level perceptual experience without antecedently determining an ontology of objects to which sensory experiences apply. The random variable approach gives us a way out of this dilemma. If our perceptual experience of the world is a measurement, then the content of that experience is just the outcome of this measurement. If the space of possible measurement outcomes is well defined, then the values taken on by the random variable which characterizes measurement outcomes are also well defined, independent of the question of what is being measured. Rather than say perception tells me of something that it is red, or that I experience redly, I prefer to say the measurement performed by my perceptual system delivered the result red.

None of this is meant to deny that we experience a world populated by objects to which we attribute properties. But treating low level perceptual states as measurements will assist us in the problem of understanding perceptual error. My low level experience of red is directly causally connected to the world and, correspondingly, it cannot be in error. However, as the low level signal is processed by the hard wiring of my visual system, the objects and properties which populate my experience are calculated from these initial measurements. This calculation may produce a model of the world which is incorrect as the result of systematic biases in the hardwiring (these systematic biases are exploited by optical illusions and other perceptual tricks). Already, then, we have a form of primitive detachability, resulting simply from staving off the attribution of properties to objects until a later stage in the perceptual processing chain.

It is important to note that the random variable paradigm is not inconsistent with the object - property paradigm, just more general. Given a set of mutually inconsistent properties (say colors), the range of a random variable may be associated with the attribution of these properties to a particular object. For example, \( X_\sigma \) might be a random variable which ranges over the possible surface colors of the object \( \sigma \); then \( X_\sigma = \text{red} \) just means that \( \sigma \) is \( \text{red} \).
1.5 Plan of the Work

The first order of business will be to develop and clarify the claim that perceptual experience exhibits a hierarchical structure. This is the purpose of Chapter 2. In order to develop this claim, we will examine some of the features of the scientific theory of color vision very closely. The purpose of such a close examination is to help show as specifically as possible just what follows from the scientific theory and what does not. For example, my claim that we can isolate within perceptual experience those features which carry the exact same information as the coding of the color signal at the retina by cone cells will not make much sense until the exact features of this coding and the corresponding experience are laid out in detail.

Of course, the other purpose of laying out the details of the scientific evidence is to demonstrate what it does not (yet?) show. For example, claims made by Paul Churchland (2005 and elsewhere) about the reduction of color experience to neural firing are not supported by the scientific evidence. This point is independent of the philosophical worries which arise if one attempts to identify experience with neural firing (though I am sensitive to these worries and feel they defeat Churchland’s too quick argument). The evidential point is that the neural structures which Churchland’s argument assumes have not been found in the brain (and the current best guess of color scientists is that they will never be found).

The remaining chapters will develop a naturalistic theory of informational content, first as applies to continuous representational spaces, then as applies to discrete representational spaces. In the final chapter, this will be developed into a tentative theory of intentional content, which will be applied to the problem of metamers in the philosophy of color. Figure 1.2 provides a rough map of the logical structure of the work.

In Chapter 3, I will develop a theory of natural measurements, of which perceptual states are an instance. I will demonstrate how this theory applies to color perception as outlined in the previous chapter. These measurements are natural in the sense that they result from causal processes (rather than the intentionality of a scientific investigator). However, they are measurements in the sense that the informational
relationship between perceptual experience and the world can be characterized in a very precise way using the theory of measurement. I will respond to one recent argument (due to Bas van Fraassen) that natural measurement is impossible.

We will also begin to address the question of error in Chapter 3. An important strategy is to distinguish several distinct types of error, which emerge at distinct points in the processing chain. The essential point here is that intentionality does not arise in one step, but gradually. Furthermore, the claim that there can be no error if we consider only causal processes will be shown to be false. At the very least, the behavior of our cone cells exhibits imprecision in the detection of light. Once we consider higher stages of the processing chain, inaccuracy is introduced in the form of systematic biases or distortions. What we do not yet have is full detachability.

Chapter 4 will investigate the transition from continuous to discrete representational spaces. The particular example discussed will be the relationship between color terms and perceptual experiences of color. The discussion will largely follow the theory developed by Gärdenfors (2000). An essential point here is that color experience derives its content from its causal relationship to the world, while color terms derive their content from the organizational structure of color experience. Once we see this distinction, we can understand how color terms can be detached from color experience.
and used to make false claims about the world.

Chapter 5 develops a general theory of the informational content of natural signs and applies this to discrete representational spaces such as those discussed in Chapter 4. The starting point of this analysis is the rejection of Dretske’s stipulation that a signal only contains the information that $p$ if the conditional probability of $p$ given the signal is 1. I will argue that this conclusion rests on a misinterpretation of the argument by Grice (1957) that natural meaning is factive. If we take the typical examples of natural signs (smoke as a sign of fire, spots as a sign of measles, etc.) as the constraints on any theory of natural content, we will see that only a probabilistic theory is adequate to explain them.

Chapter 6 will apply the analysis of the previous chapters to the problem of metamers, positing a tentative strategy for deriving intentional content from informational content. Metamers are physically distinct lights or surfaces which are perceptually indistinguishable in color. They pose a challenge for color realists as they appear to constitute categories which are not “physically interesting”. Some realists (e.g. Churchland, 2007) attempt to save metamers by providing an account in which they constitute a physically natural category; others (e.g. Byrne and Hilbert, 2003) bite the bullet and argue for realism despite the lack of physical motivation for color categories. I will argue that the realism question is simply the wrong question to ask. However, there are questions of representational efficiency and robustness of categories which appear repeatedly throughout this literature, and these can effectively be addressed by the information theoretic approach. In particular, I will investigate the possibility that the ecological account (as discussed in, say, Hatfield, 2003a) can be recovered from information theoretic analysis (without recourse to evolutionary history, the strategy usually employed in the ecological literature).
Chapter 2

The Hierarchical Structure of Color Perception

2.1 Introduction

The purpose of this chapter is twofold. First, to defend the claim that perceptual experience has a hierarchical structure. Second, to introduce the scientific evidence related to the lowest level in the visual portion of this hierarchy, what I will call the perception of colors in isolation, or *pixel color*. Along the way, we will consider the problem of how to define color.

A hierarchical structure is one which is composed of substructures arranged in levels such that higher levels of structure are composed of, derived from, or somehow supervene on lower levels of structure. At the base of this hierarchy is a lowest level of primitive structural elements. Hierarchical structures appear everywhere around us—our companies are organized by hierarchical management structure, our computer programs utilize a hierarchical software structure, biological systems (from single cells to humans) exhibit hierarchical structure, and the most popular and compelling view of the physical structure of nature itself is that more complex systems, from rocks to planets to lifeforms, are hierarchically based upon a set of low level physical primitives, whether these be atoms, quarks, or “quantum foam.” In fact, Herbert Simon (1962) has argued that hierarchical structure is a general feature of naturally
occurring complex systems.

Although it is natural to think of many of the systems we encounter in nature as hierarchically structured, the claim that perceptual experience itself is hierarchical may sound puzzling. This is perhaps because of the intuitive appeal of what Dennet (1991) has called the “Cartesian Theater”: the idea that conscious experience is a unified whole, presented in a single process, much like a movie projected on a screen. Although the critiques Dennett has launched against the Cartesian Theater view are damaging, it is difficult to produce a convincing alternative (Dennett’s own “multiple drafts” model is ultimately unsatisfying). My purpose here is not to respond to these worries by suggesting a new theory of consciousness. Rather, I merely wish to defend the descriptive claim that experience exhibits hierarchical structure and that, in at least some cases, the features of levels of this hierarchy correspond to features of a corresponding level in the hierarchical neural processing chain beginning at our perceptual organs.

The experimental methodology for investigating levels of perceptual experience is that of psychophysics. In Section 2.2 I will introduce the basics of psychophysics. I will also mention some points of contact with the current neural and philosophical literature on consciousness.

Section 2.3 will investigate the hierarchical structure of color experience in particular as it emerges from the various ways we use color terms. Two levels of this structure will be important for further discussion. The first I call the experience of colors in isolation, this is the lowest level of the hierarchy of visual experience. Color in this sense is the set of properties by which pixels in our visual field are distinguished. Under ordinary circumstances, we have only imperfect access to this level. However, the application of psychophysical methods can isolate the experience of color from other levels of our color experience. I call this level the experience of colors in isolation because color stimuli must be presented without any contextual cues (i.e. in isolation from other color stimuli) in order to produce the corresponding experience. Perhaps the most remarkable discovery of color science is that the informational relationship between isolated color stimuli and experience as revealed by psychophysical experimentation is exactly the same as the informational relationship between color
stimuli and the firing of cone cells at the retina. The purpose of Sections 2.4 and 2.5 is to lay out the evidence for this claim and to state it more precisely.

The second level of interest in the hierarchy of color experience is that associated with the perception of surface properties, namely the perception of color in context. This chapter will conclude with a brief discussion of the problems that emerge when we attempt to extend the results here to the case of color in context, attempts that have thus far failed.

2.2 Psychophysics and Beyond

2.2.1 The Science of Experience: An Example from Weber

How can we develop a scientific theory of experience? There seems to be an insurmountable problem here. Experience is inherently subjective—only I have access to my experience; yet science demands objective investigation, in the sense that any results must be accessible and repeatable by others. The “what it’s like” of a conscious experience cannot be detached from the subject who experiences it, so how can it be made accessible and repeatable to a scientific community? How can we have an objective theory of something which appears inherently subjective?

Ernst Heinrich Weber (1795–1878) developed an ingenious solution to this problem which initiated the field of psychophysics. Weber realized that although the “what it’s like” of experience cannot be directly investigated experimentally, comparisons between experiences can be investigated through repeatable experiments. In particular, we can present subjects with stimuli and ask them to report upon how they differ. For example, I may not be able to characterize scientifically what it’s like for you to experience the brightness of a light, but I can present you with two lights and ask which one appears brighter to you. This is why the field is called “psychophysics”: it investigates the relationship between physical stimuli and psychological experience. It turns out that very specific and sophisticated claims about the structure of experience can be based on experiments involving simple judgments of comparison such as this.
The basic principles of psychophysical methodology were first articulated by Gustav Theodor Fechner (1801–1887) in his seminal work, *Elemente der Psychophysik* (1860).\(^1\) This work achieved two crucial goals, first it clearly describes the basic experimental methodology behind psychophysics. We are interested in learning about the sensitivity of experience, so we would like to determine the “just noticeable difference” between stimuli which may generate an experience. We can discover this by preparing physical stimuli which form a progression (say, lights of increasing brightness, tones of increasing pitch or volume, color samples of varying hue, etc.). We then present these stimuli to the subject and ask whether he can notice a difference between each one and a comparison sample. Different methodologies present the stimuli in different orders. For example, the method of limits presents the stimuli in succession, approaching the value of the comparison stimulus. The method of constant stimuli presents the stimuli in random order. These different methodologies have practical strengths and weaknesses which Fechner discusses. His analysis was remarkably robust and these methods continue to be the standard for modern psychophysical experimentation.

The second achievement of *Elemente der Psychophysik* was its extension of the mathematical relationship between stimulus and experience discovered by Weber to a more general form. Although Fechner himself named this formula “Weber’s Law”, it is usually referred to today as “Fechner’s Law” or the “Weber-Fechner Law” in order to distinguish it from Weber’s initial formulation. Fechner’s Law is usually stated

\[
S = K \log I
\]

where \(S\) signifies sensation (the strength of the subjective experience), \(I\) signifies the intensity of the stimulus, and \(K\) is a constant, different for each domain we can sense. Essentially the law tells us that the size of a just noticeable difference in experience depends upon the absolute intensity of the comparison stimulus. For example, if the

\(^1\)It is perhaps ironic in the context of the present discussion that Fechner himself was a panpsychist. Contra the popular materialism of the day, he was a monist who thought the only substance was mind. This theoretical agenda motivated his efforts to place the scientific study of mind on a solid methodological and mathematical footing.
comparison stimulus is a very quiet sound, then I can notice the difference between its volume and that of sounds which differ only slightly in volume. If the comparison stimulus is a very loud sound, then it will take a much greater difference in volume for me to be able to hear the difference between it and the new stimulus. This law continues to be useful, although it fails to hold in the limit. A more general law replacing it was proposed by Stevens (1957 and 1961) and will be discussed briefly in the following chapter.

Let’s look more closely at an example of how a specific claim about the structure of experience can be derived from psychophysical methods. Weber discovered through experimentation that our experience of the weight of an object is actually calculated from at least three distinct perceptual faculties (perceptions of pressure, muscle activity, and temperature). As a consequence, Weber argued that the supposed “sense of touch” is really a collection of distinct sensory faculties. These each contribute differently to an experience of the weight of an object. Although we usually do not attend to these different aspects of our experience of weight, Weber was able to isolate each of them by controlling for the effects of the others. Let’s look at two examples in order to see how this works (see Weber, 1996, for details).

Experiment one. Weber was interested in the question of whether or not information from muscle activity (as when, say, one hefts an object) contributed to judgments of weight. In order to experimentally test this question, he devised a single set of stimuli and two conditions. The stimuli were just a set of small boxes, all identical in appearance and size, yet differing slightly in weight. In the first condition, subjects were asked to lay their forearms on a table (this was to prevent muscle activity from contributing to their judgments) and boxes were placed successively in their hands until the subject could detect a difference in weight. In this condition, only the sensation of pressure contributes to the experience of heaviness. Weber was able to measure the size of a difference in the weight of stimuli necessary to produce a “just noticeable difference” in the subject’s perception of heaviness. In the second condition, Weber again offered the subject boxes differing in weight until the subject was able to notice the difference. This time, however, the subject was standing up and able to heft the boxes in each hand before judging. Weber again measured the
difference in stimulus weight required to produce a just noticeable difference in the subject’s perception of heaviness. What Weber discovered was that the difference in weights required to produce a perceptual difference in heaviness was much smaller in the second condition, when subjects were allowed to use their muscles, than in the first condition. He concluded that information from muscle activity indeed contributes to judgments of weight.

Experiment two. Weber wondered whether sensibilities other than pressure contribute to judgments of weight. Weber tested whether the sensation of temperature contributes to such judgments by placing coins of different temperatures on a subject’s forehead (i.e. out of sight of the subject). Weber discovered that cold coins were judged to be much heavier than warm ones. For example, if a single silver Thaler cooled to near freezing and a stacked pair of silver Thaler at room temperature are placed on a subject’s head in alternation until the subject can form a judgment about their respective heaviness, the subject will judge them to be equally heavy, or even that the single cold Thaler is heavier than the two stacked warm Thaler. Weber concluded that sensations of temperature indeed contribute to judgments about weight.

I take these investigations to support the thesis that our perceptual experience exhibits a hierarchical structure. Sensations of temperature and pressure are more primitive in this structure than sensations of heaviness. With these experiments, Weber has demonstrated that the sensation of heaviness is derived or calculated from the more primitive sensations. He has also demonstrated, however, that we can isolate the effects of each of these primitive sensations in producing a higher level sensation of weight by simply controlling for the other sensations. Although under ordinary circumstances we cannot experience a sensation of weight calculated only from temperature and pressure, Weber was able to create a scenario in which this was the case, and thereby measure the effect of the sensation of temperature on the sensation of weight.

Finally, it is important to emphasize again what psychophysical methodology cannot reveal, the “what it’s like” of a sensation. From the standpoint of psychophysics, this is not important. What is important is that there be multiple distinct and mutually exclusive “what it’s like”s within a single sensory modality. Then, psychophysics
can tell us how these “what it’s like”s relate, which are closer to others, which are triggered by what physical stimuli, etc. An experience of heaviness by itself cannot be investigated, but as long as we can experience different degrees of heaviness, the relationship between these and to physical stimuli can be measured. Likewise with color: from the standpoint of psychophysics, the most important point about color experience is that there are different color experiences, not what any particular color experience is like.

2.2.2 Hierarchical Consciousness in Philosophy and Neuroscience

It is natural to assume that our judgments of weight are literally calculated from sensory input. Likewise for perceptual sensations in the visual and auditory domains. This is because the neural wiring of the brain is naturally interpreted as exhibiting a hierarchical organization. Input from the environment at our cone cells, for example, is processed in the retina before it is sent to the lateral geniculate nucleus for more processing, and then on to V1. Of course, the hierarchical structure here is quite complex, involving multiple parallel streams of processing. But along each stream, levels further from the retina exhibit more sophisticated and narrow responses to variations in the environment, and these responses are clearly calculated from the behavior of areas lower in the processing chain.

But what about neural theories of consciousness? Are there any that would be compatible with the claim that our conscious experience can be manipulated to correspond to different levels of this hierarchy? Perhaps the most important issue here is localization—if the neural correlates of conscious experience are restricted to a particular area of the brain, then only structure at that level of processing should be found in conscious experience. As it turns out, however, Dennett’s criticisms of a Cartesian Theater have emerged independently within the neuroscience literature. Some of the most cutting edge work on consciousness in neuroscience associates it not with a particular brain region, but with a type of neural interaction.

Victor Lamme, for example, has developed a sophisticated theory of consciousness
from the neuroscience perspective which identifies conscious experience with recurrent processing, the top down activity of neurons that can occur when a signal arrives from lower down the processing chain.

We know that neural (including cortical) activation does not necessarily lead to awareness. Hence the search for the [neural correlates of consciousness]: the investigation of what kind of neural activity is, and is not, capable of producing awareness. With respect to this question, I have made a point of abandoning the localizationist approach and instead distinguish between the so-called feedforward sweep (FFS) and recurrent processing (RP). (Lamme, 2003, 15)

This theory is supported by a number of experiments Lamme and his colleagues have performed on the effect of transcranial magnetic stimulation (TMS) on visual awareness. TMS is a technique which temporarily disrupts neural processing in a targeted location in the brain. Lamme and colleagues noticed that, although the feedforward signal from a visual stimulus arrives in V1 30–40 ms after stimulus onset, recurrent signals (feedback from later stages in the visual processing chain) only arrive at V1 around 100 ms after stimulus onset. In order to test the theory that recurrent processing is the neural correlate of consciousness, they used TMS on V1 approximately 100 ms after stimulus onset while subjects performed a visual discrimination task.

Lamme and colleagues discovered that this experimental setup induced so-called “blindsight” behavior, in which subjects perform well at a visual task but report no conscious experience of being able to see. In this case, subjects in the TMS condition performed well at the discrimination task (judging the orientation of a bar, or the color of patch, correctly 80+% of the time), but reported being unable to see the bar (or color of the patch). The explanation here is just that the signal from the forward sweep contained enough information to guide action on the task, but since the recurrent signal at 100 ms was disrupted, the subject did not have conscious access to that information.

This hypothesis states that the “feedforward sweep”, that is, the initial volley of visual information transfer through the successive visual areas, can generate motor responses but is not generating conscious experience.
Only after higher visual areas send signals back to lower areas, using recurrent processing, is an organized visual percept thought to arise. This would fit exactly with the TMS results of Boyer et al. (2005): activation via the feedforward sweep is unimpaired by the TMS pulses at 100ms latency, and could therefore allow the blindsight behaviour ... But the feedback signals arrive in V1 at exactly this latency, and will be disrupted by the electromagnetic pulses of the TMS ... Similarly, in patients, signals bypassing V1 will generate feedforward activation of extrastriate areas. Owing to the lesion, feedback signals cannot reach V1, hence there is no recurrent processing and no conscious experience. (Lamme, 2006, 194)

The theory defended by Lamme is at least consistent with the view presented here. In fact, it supports it to the extent that levels of processing to which recurrent signals are sent can potentially generate conscious experiences. What about current philosophical approaches to consciousness?

At first glance, the most relevant philosophical theories of consciousness should be those for which consciousness is a higher order phenomenon. These come in a variety of flavors, but are usually cashed out in terms of either “higher-order perceptions” (HOP) or “higher-order thoughts” (HOT). HOP theories identify consciousness with an inner sense or perception of brain states; they date back at least to Locke and have been defended more recently by Armstrong and Lycan (e.g. Armstrong, 1984; Lycan, 2004). HOT theories identify consciousness with thoughts about thoughts and are most closely associated with the work of David Rosenthal (e.g. Rosenthal, 2002).

There is danger of confusion here, however. The hierarchical structure of experience which I am defending indeed has higher and lower levels, but it is not in virtue of the higher levels themselves that conscious experience arises. I make the weaker claim that the levels of this hierarchy are accessible to conscious experience in suitable experimental circumstances (and maybe also through selective attentiveness). So, the hierarchy of perceptual experience I am describing here is not the hierarchy of higher order thoughts or perceptions in HOP and HOT theories; the two hierarchies are orthogonal.

That having been said, there is a close intuitive fit between my analysis of experience, Lamme’s analysis of the neural correlates of consciousness, and higher-order theories. The recurrent signals in Lamme’s approach are at least in a very rough
intuitive sense signals “about” the level of the feedforward chain to which they return since they are triggered by and targeted at this level of processing. In certain intuitive respects, then, recurrent signals can be interpreted as second-order. Additionally, HOP and HOT theories share with Lamme’s theory the important property of non-localization: conscious experience is not identified with a particular location, but with a type of mental activity—presumably any mental state is a potential object of higher-order thoughts or perceptions.

A potential breakdown in the analogy comes if third or more than second-order thought / perception is required. For example, Rosenthal argues that a second-order state is necessary for a conscious experience (a “what it’s like”), but a third-order state (a HOT about a HOT) is necessary in order for this “what it’s like” to itself be consciously accessible.

So if the HOT hypothesis is correct, it will rarely seem, from a first-person point of view, that HOTs accompany one’s conscious sensory states. Our first-person access reveals correlations only with conscious HOTs, not HOTs generally. And HOTs are conscious only in those rare cases in which one has a third-order thought about the HOT. But on the HOT hypothesis, HOTs need not be conscious for there to be something it’s like to be in the target sensory state. So we cannot hope to test the hypothesis by correlating in a first-person way the occurrence of HOTs with there being something it’s like to be in conscious sensory states. (Rosenthal, 2002, 413)

How does this concern fit with the results of the Lamme experiments? The suggestion I have been developing so far is that in the TMS condition, the stimulus is “sensed” in that subjects can respond to it appropriately, but there is no experience of this information due to the disruption of the recurrent signal. In the non-TMS condition, the recurrent signal is associated with an appropriate second-order phenomenon, and hence there is a conscious experience of the stimulus. However, if Rosenthal’s analysis is correct, that this is not enough to get the subject to report on that experience, but instead an appropriate third-order phenomenon is needed, then the analogy with Lamme breaks down. This is because, although it looks as if recurrent signals can, at least in an intuitive sense, be “about” a level of the feedforward processing hierarchy,
they cannot be “about” other recurrent signals in this sense. Signals are sent from neuron to neuron, they are not be sent from neuron to signal. As far as I can understand, in Lamme’s current theory, the distinction between experiencing and reporting on experience is cashed out in terms of the location of the recurrent signals. For example, if recurrent signals occur along the visual processing chain, but there is no connection to language areas in the left hemisphere, then subjects experience visual sensations without being able to report on them.

So there are areas of breakdown between higher-order philosophical theories and the Lamme theory of consciousness. Clearly, there is much to do in terms of developing a closer connection between theories from neuroscience and those from philosophy here. The main purpose of this section, however, has merely been to demonstrate that there are popular theories in both field which are consistent with the claim that perceptual experience exhibits hierarchical structure. In the following section, this claim is extended to the case of color experience in particular, which itself exhibits hierarchical structure.

### 2.3 What is Color?

If perceptual experience exhibits hierarchical structure, where does color fall within this hierarchy? The purpose of this section is to argue that color experience itself also exhibits hierarchical structure. We can see this from the way we use our color vocabulary, applying it to different aspects of experience (and correspondingly the world) depending upon the focus of our attention. Amongst the levels of our hierarchical experience of color, two are especially important for philosophical purposes. The first is what I will call *pixel color*, or the experience of colors in isolation. This is the very lowest level of the visual processing hierarchy. The second is the perception of colors as properties of surfaces. *Surface color* is perceived only in context, since our attribution of colors to surfaces depends upon contextual cues.

It is very difficult to give a non-exemplar-based definition of color. As an illustration, consider the definition of color given by the *International Lighting Vocabulary*:

> Attribute of visual perception consisting of any combination of chromatic
and achromatic content. This attribute can be described by chromatic color names such as yellow, orange, brown, red, pink, green, blue, purple, etc., or by achromatic color names such as white, gray, black, etc., and qualified by bright, dim, light, dark, etc., or by combinations of such names. (quoted in Fairchild, 2005, 84)

This provides a rough starting point, but it would be desirable to find a definition (or definitions) of color which do not depend upon exemplars.

A natural place to start in the attempt to define color is by examining our use of color terms. Some examples will illustrate the diversity and complexity of color phenomena.

1. Grass is green.
2. The sky is blue.
3. The night was inky black.
4. The ruby glinted red in the sunlight.
5. The room was suddenly flooded with white light.
6. Are you feeling alright? You’ve turned a horrid shade of green.
7. The banana is yellow, but it looks orange in this light.
8. The forest floor was a dappled grey and gold.

Sentence (1) is a typical example of the use of color terms motivating the philosopher’s question “are physical objects colored?” This is because (1) clearly attributes a color to a physical object, most obviously its surface. It is not surprising that this is taken as the prototypical use of color terminology given the etymology of the world “color”, which comes from the Indo-European root *kel* – to cover or conceal. In its suffixed form *kel-os*, it is the root for “color” and means something like “that which covers” (Watkins, 2000, 38–9).

However, to what extent does the English language category of “color” correspond to a natural categorization of surfaces? If we canvas a wide array of languages, we
discover that prima facie “color” terms of some languages encode surface properties other than those associated only with hue or chroma. For example, Hanunóo is classified by Berlin and Kay (1999) as a stage III language, with four basic color terms, corresponding roughly to black, white, red, and green (28–9). Conklin (1955) has demonstrated, however, that application of these terms varies not just with hue, but also with other parameters of surface variance in the Hanunóo environment. For example, the terms also code “an opposition between dryness or desiccation and wetness or freshness (succulence) in visible components of the natural environment” (342). As a consequence of these additional properties of surface variance coded, speakers of Hanunóo demonstrated little convergence in their application of color terms to artificial samples (e.g. prepared fabrics and Munsell color chips), but demonstrated substantial convergence when applying color terms in contrastive situations within their natural environment. The work of Berlin and Kay and the worry that cross-cultural studies like theirs are insufficiently sensitive to the social construction of our concept of color will be discussed in more detail in Chapter 4, Section 4.

Our basic application of color terms pertains not just to physical surfaces, but also apparent surfaces. Consider, for example, (2) and (3), where color terms are applied to the sky or some plane of homogeneity in the visual field. In fact, given our everyday experience with color as a surface property, it is perhaps unsurprising that ancient and medieval theories of “the vault of the heavens” treated the sky as a distant surface, rotating above us. It is only after having developed a nuanced theory of light and atmosphere that we can assert that the blue of the sky is not a surface property. Instead, it is the product of the scattering of short wavelength light from solar rays hitting the atmosphere obliquely (Nassau, 2001, 235).

(4) demonstrates we apply color terms not only to surfaces, but also solids. The redness of a ruby permeates the entire translucent whole. An interesting aspect of (4) is the term “glint.” Glinting is a type of interaction between surfaces and light not typically classified as part of color. Glinting seems to involve more than straightforward hue attribution to a surface, describing instead something about the way the color appears to emanate from the surface, or the way the light and the surface interact.
From a physical standpoint, ruby offers a good example of the difficulties which arise if we attempt to argue that color experience is correlated with physical properties of the object. The red of a ruby is caused by chromium impurities in corundum crystal. Pure corundum is inert and thus colorless; when Cr$^{3+}$ ions are embedded within the corundum lattice, however, the crystal takes on a deep red appearance. One reason for this is that the Cr$^{3+}$ ions are susceptible to two forms of energy excitation, one at 2.2 eV (absorbing light in the yellow-green part of the spectrum) and the other at 3.4 eV (absorbing light in the violet part of the spectrum). Due to this pattern of absorption, a slight amount of blue light and a large amount of red light is transmitted unhindered (Nassau, 2001, 86–9). The consequence: rubies appear red.

However, there is a second, distinct, cause of the ruby’s red appearance. As excited electrons in the Cr$^{3+}$ ions return to their ground state, they release energy at 1.79 eV, corresponding to an emission of red light—so rubies *glow* red as well! (Nassau, 2001, 89–91) The match between the wavelengths of light transmitted by a ruby and those emitted by a ruby is purely accidental. Emeralds, for example, result when chromium appears as an impurity in beryl. Like corundum, beryl is an inert crystal, composed of a lattice with similar geometry. The molecular bonds are slightly weaker in beryl than in corundum, however. As a consequence, the energy needed to excite electrons in embedded Cr$^{3+}$ ions to a higher energy state is slightly shifted, from 2.2 eV and 3.4 eV in ruby to 2 eV and 2.8 eV in emerald. Correspondingly, the shape of the Cr$^{3+}$ violet absorbtion is changed, and its yellow-green absorbtion when embedded in corundum is shifted to a red absorbtion when embedded in beryl. This allows light from the green-blue part of the spectrum to be transmitted, resulting in emerald green (Nassau, 2001, 91–3). As a final wrinkle, when the electrons in the excited Cr$^{3+}$ ions in a piece of emerald return to their ground state, they emit energy at 1.83 eV, so, just like rubies, emeralds also glow *red*!

A use of color terms taken by some philosophers to be extended, metaphorical, or, at least, irrelevant to the color realism debate, pertains to light, for example (5). The color of a light source can only be perceived directly under special circumstances, but we can infer it from the appearance of familiar surfaces, or even from simple comparison of the relative colors of surfaces falling under the light. It is of interest that
in discussing a single object, we can distinguish color in the sense of the appropriate surface property and color in the sense of a variation in appearance of that property due to variations in the light source. Frequently, we use the locution “looks $x$” (where $x$ is a color term) to distinguish this non-veridical surface property from its “true” color, as in (7). However, we can also apply color terms directly as in (8). The forest floor is “really” brown and green, it only “looks” grey and gold due to the sunlight filtering onto it through the leaves. However, because it is the grey and gold pattern that is most salient, we can attribute this directly to the floor in casual conversation.

Another flexible usage of color terms can be seen in (6), where a primary color is attributed to human skin. Other examples include the color-based racial terms which have recently fallen out of favor and locutions associated with emotions (“flushed red with anger,” “turned green with envy”). On the one hand, there is an obvious element of metaphor here: a native american’s skin is not strictly speaking red, just as a sick person’s skin is not strictly speaking green. On the other hand, however, the use of color terms in this domain is systematic. Gärdenfors has proposed an analysis of the semantics of color terms which accounts for this systematic application. He argues that if we restrict our attention to the subspace of color percepts associated with the possible hues of human skin, but apply color terms in the same relative positions on this new space as on the original space, we can produce a semantics for these examples (Gärdenfors, 2000, 121, see discussion in Chapter 4).

How are these many different types of color attribution related? The Optical Society of America (Committee on Colorimetry, 1953, Chapter 5) describes these distinct spaces of color attribution as “modes” of color viewing. They identify five distinct modes:

1. Illuminant
2. Illumination
3. Surface
4. Volume
5. Film
(3) and (4) should be obvious: the attribution of color to surfaces and translucent volumes. (1) and (2) are distinguished in that (1) occurs when we attribute color to an object, a glowing or emitting source (“The flame of the bunsen burner glowed blue in the dusky lab”); (2) occurs when we attribute color to ambient lighting, without identifying a source or particular glowing object (e.g. “the light at dusk turned an eerie pink”).

The unfortunately named “film” mode occurs when colors are viewed through an aperture with no connection to an object. These are color perceptions produced in isolation from any context (including additional features of the object / surface itself which might influence the color judgment). Crucially, “[a]ll other modes of appearance can be reduced to film mode” (Fairchild, 2005, 144; Committee on Colorimetry, 1953, 146). This is because the film mode isolates color percepts, and the appearance of color in a complex scene is derived from the appearances of colors in isolation.

The “film” mode of color appearance, colors as they appear isolated from contextual cues, constitutes the lowest level in the hierarchy of visual experience, as well as in the hierarchy of color experience. The purpose of the following two sections is to demonstrate the close structural correspondence between properties of the “film” mode of color perception and the properties of the earliest stage of visual processing, the encoding of information by cone cells in the retina. This close correspondence motivates the notion of pixel color:

**pixel color** – that property by which pixels in the visual field are categorized and distinguished.

Pixels are just the discrete uniform regions which via their arrangement make up an overall image, as on a TV screen or in a photo. The retina registers incoming light via the activation of discrete cells, which form the “pixels” of the encoded image. So, this name is motivated by the physiological structure to which this level of perceptual experience (the OSA’s “film” mode, Fairchild’s “unrelated color” (Fairchild, 2005, 145)) corresponds.

The second sense of color which will appear frequently in following chapters is that which has historically been of greatest interest to philosophers:
**surface color** – that property by which surfaces in the visual field are categorized and distinguished.

The sense in which surface color constitutes a higher level in the color experience hierarchy is twofold. First, on the physiological side, it corresponds to structures much further along the processing chain than those to which pixel color corresponds. On the side of experience, all plausible models of the appearance of colors in context are calculated from the physical stimulus arriving from the surface *plus* other features of the scene, whether it be average luminance levels, the physical stimulus arriving from a neutral white surface, whatever. This is the sense in which other modes of color viewing “reduce” to “film” mode: the “film” mode involves a single color signal, but other modes must be calculated from multiple color signals.

One way to make sense of sentences such as (7) and (8) is that attention (and correspondingly application of color terms) is shifted from surface color to pixel color. Although the surface color of the forest floor is brown, I attempt to describe my experience of it in more detail by shifting my attention to the color experience associated with uniform regions in my visual field (i.e. pixel color). I notice that these are grey and gold. This explains how a surface can be “both” *brown* and *dappled grey and gold*—at one level of the perceptual hierarchy it is brown, at another it is grey and gold. It is important to note, however, that complete recovery of the properties of pixel color is not in general available when viewing a complex scene. This is why the detailed exploration of this low level in the perceptual hierarchy requires special experimental conditions.

In the following two sections we discuss in detail the evidence for a close correspondence between the coding of the color signal at the retina and pixel color perception. The chapter concludes with a brief discussion of the difficulties involved in attempting to extend this detailed correspondence to the case of surface color.
2.4 Transduction and Scotopic Vision

Before examining the relationship between color vision and pixel color experience, it will be helpful to first examine the simpler case of scotopic vision. Viewing conditions are said to be scotopic if light levels are so low, only rod cells are activated. If light levels are high enough to activate cone cells, the conditions are said to be photopic. The scotopic case is simpler for two reasons. First, there is only one type of cell activated at the retina (rather than three in photopic conditions). Second, our visual experience in scotopic conditions is monochromatic, and thus some of the complications associated with full blown color vision are avoided.

Before investigating the details of scotopic viewing, however, it will be helpful to make some general remarks about transduction. This concept is fundamental for understanding how information flows in nature and, specifically, how our perceptual system causally interacts with the world. Several of the issues raised here will appear again and again in following chapters. In particular, the fundamentally statistical character of transduction will motivate the discussion of the precision of natural measurements in Chapter 3 and the probabilistic account of informational content in Chapter 5. After the basic idea of transduction is on the table, we can discuss the features of transduction in scotopic conditions.

Section 2.5 we will extend this discussion to the case of photopic viewing conditions, when cone cells are active.

2.4.1 Transduction and Information

The most fundamental type of process for propagating information in nature is transduction.

transduction – a process by which energy is converted from one form to another

For example, whenever the human perceptual system interacts with the external world, it must first convert the energy impinging on it into the electrical activity of nerve cells. So, mean molecular motion (heat), vibrations in air (sound), photons of
particular wavelengths (light), all these different forms of energy are converted into the same type of energy by the perceptual system, the electrical activity of nerve cells. This fundamental insight is generally attributed to Johannes Müller (1842):

**the doctrine of specific nerve energies** – perception depends upon the energy reaching the brain from specific nerves. The quality of the perception depends not upon differences inherent in the nerve cells (i.e. not upon different types of energy reaching the brain), but upon the wiring of the nerve cells (i.e. which specific cells connect with which specific regions of the brain).

So, the different perceptual experiences associated with, say, tastes and sounds do not depend upon a qualitative difference in the behavior of neurons in the ear and neurons in the tongue. Instead, the differences depend upon the wiring of the neural system: neurons in the tongue and neurons in the ear are connected to different regions in the brain. It is this difference in wiring which creates perceptual differences.

Although transduction is a crucial concept for understanding how perception works, it can also be applied to human-independent informational relationships in nature. For example, the interaction between light and matter which results in the appearance of color is also a variety of transduction—more specifically, a number of distinct types of transduction! Consider the example of rubies, discussed above. The \( \text{Cr}^{3+} \) ion embedded in a lattice of corundum absorbs photons of a specific wavelength because they have the appropriate amount of energy for shifting its electrons to the next valence level. So, some of the energy in the light is transformed into energy in the ion as it enters an excited state. The red glow a ruby emits also involves transduction; the energy of the excited electrons is transformed into energy in the form of photons as the electrons return to their ground states. These are only two examples of the many distinct types of transduction which can occur when light interacts with matter.

Many instances of transduction are inherently statistical in character. The excitation of the chromium ion is a sign of the presence of photons of a certain wavelength. This relationship is not absolute, however. On the one hand, there may be photons of that wavelength which pass through the ruby undisturbed, passing too far from
any chromium ion to be absorbed. Conversely, there is really a range (albeit narrow) of wavelengths which will excite the Chromium ion; if we observe it in a state of excitation, we cannot infer that a photon of this precise wavelength was absorbed, but rather that a photon of this wavelength was absorbed with probability $x$ and one of that wavelength absorbed with probability $y$. This same probabilistic character can be observed in the receptivity of the cells of the retina.

### 2.4.2 Transduction in Rod Cells

When light passes through the lens of the eye, it is focussed onto the the retina, the back wall of the eye ball. Transduction occurs in four types of cells arrayed across the retina: the rod cells and three types of cone cells. Rod cells produce the achromatic vision we have under very low lighting conditions. It will be useful to examine the process by which transduction occurs in rod cells in detail before moving on to a discussion of cone cells.

Suppose you are in a pitch dark environment, your eyes are open, and adapted to the dark, but still you see nothing. Now suppose a stray photon passes through the lens of your eye—what effect on your vision does this photon have? The first, perhaps obvious, point is that photons are very tiny. It is possible, then, that the photon passes through the retina without traveling near enough to any rod cells to be absorbed (and presumably to be absorbed shortly thereafter by some tissue or other; our heads are, after all, opaque). If, however, the photon does pass near enough to a rod cell to be absorbed, then the rod cell will fire. That’s right, it only takes the absorption of a single photon to excite a rod cell! This brings us to one of the fundamental concepts for understanding transduction in the eye, the *principle of univariance* (Rushton, 1965).

**principle of univariance** – “the absorption of a photon leads to the same neural response, no matter what the wavelength of the photon” (Wandell, 1995, 92)

It is the principle of univariance which ensures that the informational content of rod firing is inherently stochastic. Before looking at what happens after the rod has fired,
let’s examine the process of absorption in a little more detail.

Rod cells contain a special molecule called rhodopsin. Rhodopsin is a photopigment, a molecule which is sensitive to light. When rhodopsin absorbs a photon, it changes shape. In the dark, rhodopsin is curled into a moderately unstable configuration. Once the molecule is excited by a photon, it transitions from this state into a straight configuration, decomposing into opsin and vitamin A (Wald, 1967). It is this change in the structure of the rhodopsin which triggers the firing of the rod cell. Once it has decomposed, the rhodopsin is said to be bleached. Before the rhodopsin molecule can participate in registering the presence of light again, it must return to its original configuration. This cycle of changes in the molecular structure of the photopigment occurs in the exact same fashion when the rhodopsin is excited regardless of how exactly it is excited, hence the principle of univariance (see Figure 2.1).

Although the rhodopsin’s response is binary (it decomposes or it does not), it’s receptivity to photons is statistical in character. We can see this by exposing a large number of rhodopsin molecules to different wavelengths of light at the same intensity. We discover that the sensitivity of rhodopsin peaks around 500 nm (see the black circles in Figure 2.2). In the case of a single molecule of rhodopsin, this just means it is more likely to absorb a photon with wavelength 500 nm than one with 550 or 450 nm. In the case of many rhodopsin molecules, this means that more of them will react when exposed to 500 nm light than when exposed to 550 or 450 nm light. A propensity

Figure 2.1: Sequence of structural changes as rhodopsin absorbs light. (Wald, 1967, Figure 7 (302))
Due to the principle of univariance, an identical response in rhodopsin can be achieved by any wavelength of light within the range 400 - 600 nm, so long as the intensity of the light is increased appropriately. For example, in Figure 2.2 we see that the sensitivity of rhodopsin to light at wavelength 545 nm is about half that of its sensitivity to light at wavelength 510 nm. If we expose some rhodopsin to 510 nm light at intensity $I$ and discover that $x\%$ of the rhodopsin reacts, then we can achieve exactly the same effect ($x\%$ of the rhodopsin reacts) by simply exposing the rhodopsin to 545 nm light at intensity $2I$.

Only a single photon is needed to initiate the decomposition of rhodopsin. Although each rod cell contains many molecules of rhodopsin, only a single molecule needs to decompose for the rod to react. How does this reaction translate into a perceptual experience? Can we “see” a single photon? The answer is no. Rod cells are not wired directly to visual regions in the brain. The rod signal is mediated by
several layers of processing in the retina itself. The signals from a group of nearby rod cells converge on a single intermediary neuron for comparison before it sends a signal to the brain.

So how many photons of light need to be absorbed before we have a perceptual experience? Hecht et al. (1942) constructed an apparatus for flashing a very small number of photons at a subject. They used photons of wavelength 510 nm and a combination of filters and apertures to ensure control over the number of photons incident on the eye in a single flash. The minimum number of photons needed for a perceptual experience to register ranged from 54 to 128. However, this is not indicative of the number of photons absorbed by rhodopsin in the rods. Hecht et al. calculate that approximately 50% of the photons are absorbed by the ocular media between the cornea and the retina. Then, most of the photons pass through the retina itself unabsorbed due to the sparsity of rhodopsin. Hecht et al.’s final assessment (the standard number today) is that approximately seven photons need to be absorbed in order for a perceptual experience to occur.

Hecht et al. (1942) used photons at a single wavelength, 510 nm, to which the rhodopsin would be maximally sensitive and thus, presumably, a perceptual experience most likely to occur. Can we say more than this? How close is the relationship between the absorption properties of rhodopsin and our perceptual experience in low light conditions? It turns out that the answer is extremely close. Consider this psychophysical experiment: we ask the subject to look through a hole into a divided room. We control the lighting on one half of his visual field, he controls the lighting on the other half. The room is constructed so that there are no visual cues or variations in its surfaces, all the subject sees is some homogeneous brightness on each side.

Now, call the light that we control the test light and call the light the subject controls the primary light. The primary light has a wavelength of 510 nm and the subject is able to control its intensity with a single knob. Now, for the test lights, we present the subject with a succession of lights of different wavelengths, all with exactly the same intensity. We ask the subject to adjust the intensity of his primary light so that the two lights match exactly. For example, if we present the subject
with light at intensity $I$ and wavelength 545 nm, he may adjust his primary light to an intensity of $I/2$. This tells us that the subject perceives 545 nm light as half as bright as 510 nm light at the same intensity. Figure 2.2 graphs the outcome of such an experiment against the outcome of the sensitivity experiments on rhodopsin. Note the close match between the white circles (perceptual sensitivity) and the black circles (rhodopsin sensitivity).

What can we conclude about pixel color perception under scotopic conditions? First, a key property of the perceptual experience of brightness for isolated stimuli, namely sensitivity to wavelength, matches the behavior of this same key property at the level of photopigment absorption. This speaks strongly in favor of the idea that, when we perform a light matching experiment such as the one described above, where there are no contextual cues for assessing brightness, we can isolate that aspect of perceptual experience which matches the behavior of rods in the retina. However, due to the principle of univariance, we cannot say for any particular perception of brightness what wavelength of light produced it. At most we can only recover information about the incoming signal relative to some fixed standard. Even though we receive no absolute information about the incoming signal, it is clear that as properties of the light vary, the perceptual experience varies systematically. For example, if the wavelength is held fixed, and the intensity increased, then the perceptual experience of brightness will increase. If the intensity is held fixed and the wavelength shifted, then the perceptual experience of brightness will increase as the wavelength approaches 510 nm in accordance with Figure 2.2.

2.5 Pixel Color Perception in Photopic Conditions

As the scene brightens, we shift from scotopic to photopic viewing conditions, and the cone cells become active. Where scotopic viewing conditions produced only monochromatic perceptual experiences, photopic viewing conditions produce chromatic experiences due to the three distinct types of cone cell. If we study the perception of colors in isolation, can we find the same convergence in properties of the physiological and perceptual aspects of color vision as we found under scotopic conditions? The answer
is yes; however, in order to make sense of the convergence, we must think in terms of coding and information. The response of cone cells does not map directly to perceptual experience. However, if we think of the cone cells as encoding information in the light incident on the retina, we see that the perceptual experience contains the same information as is encoded by the cones. So, even if perceptual experience of colors in isolation depends upon some cortical representation which differs qualitatively from the representation at the level of the retina, both representations contain exactly the same information about the incoming light.

2.5.1 Transduction in Cone Cells

In a normal observer there are three types of cone cells, commonly called L, M, and S cones (for long, medium, and short wavelength sensitivity). Although the cone and rod cells constitute the “pixels” which sample light incident on the retina, it will perhaps be helpful to point out some differences between cone cells and the pixels one is used to from TV screens and digital cameras. Most importantly, a cone cell does not, by itself, produce any kind of color signal. It is only by combining the signals from the three different kinds of cone cells that color information is produced. This is why vision under scotopic conditions, when only rod cells are active, is achromatic. Cone cells, just like rod cells, are subject to the principle of univariance, thus their behavior does not tell us directly what the wavelength of the exciting photon(s) is. It is the different patterns of sensitivity exhibited by the three types of cones which allows for chromatic perception.

A perhaps counterintuitive feature of the distribution of L, M, and S cones across the retina is that they occur in very different densities. The S cones, for example, are much fewer and more sparsely placed than the L and M cones. Furthermore, their relative frequencies differ in different regions of the retina. One possible explanation for this (Wandell, 1995, 53–4) is that the (relatively) short wavelengths to which S cones are most receptive are those most distorted by the optical properties of the lens. This explanation depends upon the assumption that “visual acuity is the main factor governing the photoreceptor mosaic” (Wandell, 1995, 54).
So, the arrangement of cones is geared first towards producing a *sharp* image. The production of accurate color information is secondary to this goal. This prioritization is mirrored in one effective strategy for compressing information in digital video. Frequently, digital video samples luma (brightness) at a higher frequency than it samples chroma (color). This means that the color pixels (patches of uniform chroma) are coarser than the luma pixels (patches of uniform brightness). NTSC digital video, for example, samples brightness for every pixel on a scan line, but color only for every four pixels (this is called 4:1:1 compression). Surprisingly, it can be quite difficult to perceive any difference between two images at the same resolution when one has uncompressed color (4:4:4) and the other has compressed color (e.g. 4:2:2 or 4:1:1). The efficacy of this method of image compression is due to the visual system’s prioritization of brightness information over color information.

The exact chemical properties of the cone photopigments are not as well understood as those of rhodopsin. However, experiments can be performed that bypass the photopigment response and directly measure the cone response itself. This is much easier in the case of cones than rods because cone cells are much larger. Baylor et al. (1987) performed an experiment to measure the responsiveness of cone cells in the retina of the macaque monkey. They first removed a section of the monkey’s retina and transferred it to a petri dish, where they were able to keep the cells alive. They then sucked individual cone cells into a pipette, specially designed to allow them to flash the cell with different wavelengths of light. As they flashed the cell, they measured its electrical response.

Neurons usually have a baseline state of electrical activity. In the case of cone cells, Baylor et al. (1987) used the activity of the cell when dark adapted as a baseline. After flashing the cell with light, they measured the cell’s activation relative to this baseline (the zero point in Figure 2.3). These experiments confirmed that the univariance of photopigment response observed by Wald extends to a univariance in cone response. For example, the solid line in Figure 2.3 represents the response of a macaque cone cell when flashed with light at 500 nm. The dotted line represents the same cell’s response when flashed with light at 659 nm, but at 9.3 times the intensity of the 500 nm stimulus. Notice that the general pattern of the cone cell’s response is the same
As was expected, Baylor et al. discovered three distinct types of cone cells, each with its own distinct pattern of response. The response functions are graphed in Figure 2.4. Notice that the units on the y-axis are logarithmic; so, cone responses were measured over six orders of magnitude, a factor of one million. Also notice that the response functions of the L and M cones peak very close together. This implies that (with the appropriate processing to disentangle the signals) more fine-grained distinctions can be made in this part of the spectrum.

What informational benefit comes with having three distinct types of receptors? In the case of scotopic viewing, with only one type of receptor, we saw that rod firing only distinguishes between wavelengths of light when intensity is held constant. This is just as true if we examine a single type of cone cell in isolation. With three types of cone cells, however, the wavelength of light can be distinguished by the difference in the responses of each receptor. Although 500 nm light at intensity $I$ and 659 nm light at intensity $9.3I$ produce precisely the same response in one type of cone cell, they will not, in general, do so for the other two types. So, while the response of each cone cell follows the principle of univariance, their responses can be combined to compute
Figure 2.4: Sensitivity functions for (from left to right) S, M, and L cones in the macaque retina (Baylor et al., 1987, Figure 3.A (151), reproduced by permission of John Wiley & Sons, Ltd.)
information about the wavelength of the light incident on the retina.

But the information that can be extracted from light by three univariant receptors is severely limited. Each cone cell exhibits a range of possible electrical activity. This range of activation constitutes a single dimension along which measurement can take place. Since the receptivity of each type of cone cell is different, appropriate processing should be able to produce a three dimensional representation of the light incident on the retina. But consider now the light itself—how many dimensions are needed to define the physical properties of a shaft of light? So far, we have only considered light made up of a single wavelength, which can vary in intensity. This is an artificial scenario found only in the lab, however. In the real world, light is usually composed of photons of many different wavelengths. We can represent light, then, by its spectral power distribution. The spectral power distribution of a light tells us the relative intensity of each wavelength in the light. However, even in the visual range, 380 – 750 nm, there are a continuous, i.e. infinite, number of distinct possible wavelengths. So, the space of possible spectral power distributions is infinite-dimensional. The coding of light into the S, M, and L cone signals constitutes a reduction from an infinite-dimensional to a three-dimensional space of possibilities—with corresponding loss of information! (c.f. Suppes et al., 1989, Chapter 15)

2.5.2 Color Matching Experiments

The scotopic light matching experiment described in section 2.4.2 has a more sophisticated form for photopic matching. This experiment isolates the level of pixel color space in the perceptual hierarchy in order to allow its investigation by psychophysical methods. As before, the subject views a room with no visual or contextual cues. His visual field is split by a divider in the middle of the room. On one side the experimenter shines a test light. Now, however, we give the subject access to three different primary lights (see Figure 2.5).

In the scotopic case, all lights involved comprised photons of a single wavelength. Now, however, we allow for the test light to have any spectral power distribution.

\(^2\)The presentation in this section closely follows that of Wandell (1995), 80-5.
whatsoever. So, as discussed in the previous section, the space of possible test lights is (in theory, at least) infinite-dimensional. For the moment, let us restrict the primary lights to single wavelengths. We will discuss restrictions on what primary lights can be used in a moment; for now, however, think of them as red, green, and blue lights (for example, Stiles and Burch (1959) used lights at 645.2 nm (∼ red), 523.3 nm (∼ green), and 444.4 nm (∼ blue)). After the experimenter picks a test light, the subject is asked to adjust the primary lights until the two halves of his visual field match. He can separately control the intensity of each of the three primary lights.

Remarkably, for every possible test light, the subject is able to find a combination of the three primaries which produces an exact perceptual match. In mathematical terms, any test light can be represented by a linear combination of the three primary lights. Furthermore, these representations are additive. Let $T_1$ and $T_2$ represent the spectral power distribution of two test lights and let $p_1$, $p_2$, and $p_3$ represent the three primaries. Now, to say that $T_1$ can be matched by a linear combination of the three primaries is just to say that

$$T_1 = J_1 p_1 + K_1 p_2 + L_1 p_3$$

for constants $J_1$, $K_1$, and $L_1$. Likewise,

$$T_2 = J_2 p_1 + K_2 p_2 + L_2 p_3$$
The equal sign here just means “perceptually indistinguishable,” and the addition signs signify straightforward superposition of lights (simply shining them simultaneously onto the same white surface). Since these representations are additive, we can predict the perceptual match for \( T_1 + T_2 \) by simply adding the respective constants together. In other words,

\[
T_1 + T_2 = (J_1 + J_2)p_1 + (K_1 + K_2)p_2 + (L_1 + L_2)p_3
\]

Now, some caveats. First, I claimed that every possible test light can be matched by a combination of the three primaries. This is true: every possible test light can be matched by a linear combination of the three primary lights. However, the experimental setup as depicted in figure 2.5 will not let us produce every linear combination of primaries. The setup as described above will only allow us to combine positive quantities of each primary. In order to match every possible test light, however, we will sometimes need to subtract one or two of the primaries from the other one(s). Mathematically, it is trivial to represent this possibility. In an experimental setup, however, the equivalent of subtraction can only be achieved by moving one or two of the primary lights to the opposite side of the visual field, a somewhat awkward rearrangement (though one which has certainly been performed, e.g. Stiles and Burch (1959)).

Second, what constrains our choice of primary lights? Actually, they need not be restricted to single wavelength lights, nor need they be restricted to red, green, and blue. In principle, the only restriction is that the three lights be independent, i.e. that no one light can be matched by a linear combination of the other two. This should make intuitive sense: if one light can be derived from the other two, then it does not contribute a new dimension of variance to the representational space.

So, since a match can be produced between any test light and a linear combination of any three independent primaries, there is no way that the color matching experiment alone can help us predict the physiological response of the cone cells. Furthermore, the characterization of perceptual space produced by the experiment is not
unique. One way to think about it is this: the choice of primaries determines a coordinatization of perceptual space, or, more properly, a set of basis vectors. Since the choice of vectors is arbitrary (modulo independence), we can systematically translate from any adequate coordinate system into any other. Mathematically, this is because any two sets of independent primaries can be related by a linear transformation. This means that for any two sets of primaries \( p_1, p_2, \) and \( p_3 \) and \( p'_1, p'_2, \) and \( p'_3 \) there exists a \( 3 \times 3 \) matrix \( A \) such that

\[
\begin{align*}
p'_1 &= a_{11}p_1 + a_{12}p_2 + a_{13}p_3 \\
p'_2 &= a_{21}p_1 + a_{22}p_2 + a_{23}p_3 \\
p'_3 &= a_{31}p_1 + a_{32}p_2 + a_{33}p_3
\end{align*}
\]

and for any test light \( T \), this matrix can be used to transform the representation of \( T \) in terms of primaries \( p_1, p_2, \) and \( p_3 \) into its representation in terms of primaries \( p'_1, p'_2, \) and \( p'_3 \). So, the representation of perceptual space determined by color matching experiments is unique up to linear transformation. The coordinates which characterize the point within a standard version of this space onto which a particular spectral power distribution maps are called its tristimulus coordinates.

So, now that we’ve measured both the sensitivity of cone cells and the psychophysical coding of spectral power distributions into three dimensions, can we find a linear transformation which will compare the responsiveness of the three cone cells with the tristimulus coordinates for each distinct wavelength of light? This was first done by Baylor et al. (1987), and the close match they discovered between cone responsiveness and behavior on color matching experiments can be seen in Figure 2.6.

Finally, we can recover some of the qualitative features of perceptual experience by suitably transforming our three vector representation of color matching behavior. Suppose our three basis vectors are \( R, G, \) and \( B \). If we are interested in hue rather than hue + brightness, we might consider a plane through this three dimensional space, for example the plane which satisfies \( R + G + B = 1 \). This plane is interpreted as a plane of constant brightness, capturing only hue information. A number of representations along these lines have been developed. One of the most useful is the
Figure 2.6: Comparison between values of the test lights in a color matching experiment and responsiveness of macaque cone cells. The x-axis indicates the wavelength of the monochromatic test light. The three solid lines represent the relative weights of primary lights at wavelengths 645.2 nm, 523.3 nm, and 444.4 nm (data from Stiles and Burch (1959)). The triangles, squares, and circles represent measurements taken from macaque S, M, and L cones, respectively. In addition to scaling, corrections were made for the absorption of the lens. (Baylor, 1987, Figure 16 (46), reprinted by permission, copyright (1987) Association for Research in Vision and Ophthalmology).
Figure 2.7: CIE 1931 color space, with the rough location of primary color percepts due to single wavelength light notated around the edge.

*CIE 1931* \((x, y)\) chromaticity space (Figure 2.7), defined to represent color matching behavior for an imaginary “standard observer” by the *Commission internationale de l’éclairage*. The result is a horseshoe shaped space with responses to single wavelength stimuli falling on the edge. The two ends of the horseshoe are joined by a line representing purple, which is experienced with stimuli falling at either edge of the visible portion of the spectrum. (I have suppressed many of the details of how this space is actually calculated.)

### 2.6 Conclusions and Extensions

In the next chapter, I will argue that the relationship between a space of perceptual experience, such as that of pixel color, and a space of variation in the world, in this case spectral power distributions of light, is best understood as a measurement. Before moving on, it will be helpful to conclude with some remarks about what exactly these results show and the extent to which they can be extended to our perception
of surface colors.

2.6.1 What Figure 2.6 Does and Does Not Show about Pixel Color Perception

What exactly does the argument in the preceding section, summarized in Figure 2.6, tell us about our experience of pixel color? I claim that it shows us that a domain of perceptual experience (that isolated by color matching experiments) contains exactly the same information about light impinging on the retina as the physiological response of cone cells at the retina. This implies that in whatever causal process leads from cone firing to conscious color perception, zero information is lost.

We can see that exactly the same information is contained in the perceptual experience as in the physiological response in two steps. First, the information in pixel color perception is invariant under linear transformations of basis vectors. This is experimentally verified by the fact that any three independent lights can be used to achieve perfect performance in a color matching experiment. The success of predictions of matching behavior based on the additivity of these representations demonstrates a robust understanding of the properties of these vector spaces. Second, the representation of the cone signals at the retina happens to fall within this space of equivalent representations since it also requires three vectors which satisfy the same independence conditions. Therefore all representations of the perceptual experience are informationally equivalent to the physiological representation.

This precise match between experience and physiology is truly remarkable, but it is important to emphasize what is not shown here. One crucial issue is this: the three vector representations given here are a far cry from a true color appearance model, even in the case of pixel color. A color appearance model should provide a metric over the space of possible color experiences. This means that for any three color percepts, we should be able to answer the question: is percept 2 more similar (“closer”) to percept 1 or to percept 3?

The three vector representation derived from color matching experiments does not accurately portray perceptual distances as can be demonstrated with a modified
version of the color matching experiment. Instead of asking subjects to match the
test light with their primaries, they are asked to produce a light which is the same
brightness as the test light, but a single step removed from it in chromaticity. The
exact size of the chromatic step is not relevant here, but only that the subject attempts
to produce a color the same distance from each of many different test lights. Such an
experiment was performed by W. D. Wright in 1941 and Wyszecki and Stiles have
plotted some of his data on the CIE 1931 color space (Wyszecki and Stiles, 1982,
575–6). This plot is reproduced as Figure 2.8. The dramatic difference in line length
depending upon orientation and locus within the CIE 1931 color space indicates that
the natural metric on this space does not represent the perceptually experienced
differences between color percepts.

Color appearance models attempt to characterize the space of possible color per-
cepts as a three dimensional solid with a well defined metric. In addition to char-
acterizing the informational relationship between color percepts and stimuli, color
appearance models must also account for intrinsic features of our color experience.
For example, blue and yellow (and red and green) are opponent colors, “opposites”
in the sense that we cannot perceive mixtures between them (bluish-yellow is not
a possible color). Opponent color phenomena cannot be predicted from tristimulus
coordinates, yet they place a constraint on any plausible color appearance model.

Although there are some very effective color appearance models, there is no
widespread agreement on how precisely the space of color appearances should be
characterized. Even though these spaces are often three dimensional, these three di-
mensios are qualitatively very different from the three dimensional representation
described in the above section. Popular early examples took the shape of a deformed
“spindle”, such as that in Kirschmann (1895) (Figure 2.9). In such models, the ver-
tical axis represents lightness, distance from the center represents saturation, while
the hue dimension is circular (characterizing the circular pattern of color transitions
familiar from the color wheel). More modern color appearance models use similar
axes, yet the shape of the solid is much more distorted and irregular (to conform
to psychophysical data). A historical overview of color models such as these can be
found in Kuehni (2003).
Figure 2.8: Data from Wright (1941) plotted in the CIE 1931 color space. Open circles joined by a line indicate the locus of lights judged to be a constant chromatic distance apart. (Wyszecki and Stiles, 1982, Figure 1(7.10.5), reproduced by permission of John Wiley & Sons, Ltd.)
Given these discrepancies between plausible color appearance models and the three vector representation derived from matching behavior, the result discussed above does not open the door for reductionism about color perception. Churchland (2005), for example, argues for a reduction between a vanilla version of the standard color spindle and a neural wiring model of opponent color vision along the lines of that offered in Hurvich (1981), Chapter 11. One problem here is that the specifics of the color appearance model he appeals to have not been specified. A bigger problem, however, is that the neural correlates of the model suggested by Hurvich have simply not been found. Churchland is confused here about the nature of the model; he treats it as if it is derived from our empirical understanding of neural wiring. But this is not the case. The model was originally proposed in the 1960s by Hurvich and Jameson as a potential pattern of neural wiring which might explain the qualitative features of color perception. This inspired a research program to find corresponding neural structure. But such structure simply has not been found. Derrington et al. (1984),...
for example, found neurons in the macaque LGN which respond as if receiving a red-green opponent color signal yet which “do not meet the specification of the canonical R-G opponent cell”—i.e. the cells posited in the Hurvich-Jameson model (264). As Wandell discusses:

The spectral responses of these neural populations suggest that there is only a loose connection between the signals coded by the neurons and the perceptual coding into opponent hues; it is unlikely that excitation and inhibition causes our perception of red-green and blue-yellow as opponent colors. One difficulty is the imperfect correspondence between the neural responses and the hue-cancellation measurements. The second difficulty is that there is no substantial population of neurons representing a white-black signal. This is a very important perceptual dimension which must be carried in the signals of the LGN. Yet, no clearly identified group of neurons can be assigned this role. (Wandell, 1995, 324)

This is why it is important to emphasize that only a match between the informational properties of coding at the retina and perceptual experience has been demonstrated. There are secondary properties of perceptual experience, which do not carry information about the interaction at the retina, for which neural correlates have not been found. A possibility which remains open, however, is that these secondary properties of perceptual experience (e.g. opponent color phenomena) do carry information about more distal steps in the causal chain leading from world to experience. This possibility will be explored in more detail in later chapters.

2.6.2 Extending these Results to Surface Color Perception

The philosophical debate about color has traditionally been focused on surface color perception. However, the very robust results discussed here for pixel color perception do not extend in any immediate way to the case of surface color perception.

The primary issue in investigating surface color perception is color constancy. Color constancy occurs when a surface is judged to be the same color across changes in illumination. Such judgments are surprising because the spectral power distribution of light reaching the eye from a surface will in general differ with the illumination. This is why surface color perception must lie at a higher level in both the perceptual
hierarchy and the neural computational hierarchy. To the extent that color constancy occurs, it must be calculated from the initial color signal.

However, there is already confusion about the extent to which color constancy does or does not occur. It obviously occurs because we judge apples to be red in both bright sunlight and shady fluorescent light. It obviously does not occur because we frequently differ in our assessment of color attribution depending upon lighting conditions (“in this light my pants look green, in that light they look brown”), and we can consistently apply our basic color terms to these situations in a straightforward way (see the discussion of sentences (7) and (8) above).

The depth of these worries emerges in the literature on color constancy experiments. These experiments often attempt to extend the paradigm of the color matching experiment to the matching of surface colors across changes in illuminant. One strategy uses a split screen monitor. The same scene is shown on each half of the screen under different illuminants (calculated from computer models, of course). The subject is asked to match the color of a surface on one side of the screen to the color of a test surface on the other side of the screen. He is given control over the appearance of this patch via chromaticity controls, much like the primary light controls in the original color matching experiment.

However, as discussed in Wright (2009a), subjects are often confused by the task. Their responses depend critically on the wording in the task description. Their behavior approximates that intended by the experimenter only after training. Even then, their performance does not conform to the value for the matched surface calculated from computer models. Finally, subjects themselves frequently report dissatisfaction with their match, judging it to be only approximate, or that a perfect match is impossible. Consequently, no consensus has emerged from these asymmetrical color matching experiments about how to improve models of surface color appearance.

Remember the case of the Hanunoo discussed above: their failure to consistently apply “color” terms to artificially prepared color samples inspired the worry that color as discussed in Western languages may actually not constitute a natural categorization of surfaces. Another possibility is that surface color in the sense defined above (the property by which surfaces are categorized and distinguished in experience) requires
more than the three dimensions which seem enough to characterize the perception of colors in isolation.

For example, consider the property of glossiness, the tendency to produce specular highlights. Do we ever judge of a matte surface and a glossy surface that they are exactly the same color? If not, then perhaps surface reflectance properties other than just the percent at which each wavelength of light is reflected need to be taken into account in our physical model of the relevant stimuli. Correspondingly, our perceptual model may require an additional dimension for glossiness. Similar attempts to expand the dimensionality of surface color models are presently being explored in some psychophysical experiments on color constancy.

In order to extend the robust results achieved in the case of pixel color perception to the case of surface color perception, we may need to adjust our physical models of surfaces, and we certainly will need to continue to adjust and improve our perceptual models. It may turn out that surface color demands a higher dimensional model of experience than that which works for pixel color.
Chapter 3

Perception as Natural Measurement

3.1 Introduction

In the last chapter, we examined the relationship between physical stimuli impinging on the eye and the perceptual experience generated by these stimuli. Does a perceptual experience of pixel color represent the physical stimulus which caused it in any interesting sense? The purpose of the present chapter is to argue that perceptual experiences represent features of the world in precisely the same way the outcome of a measurement operation represents the quantity measured.

In the most typical cases of measurement (e.g. of lengths, temperatures, weights, etc.), both the representing space and the represented space are continuous. In these examples, the representing space is that of real numbers, while the represented space is some domain of continuous variation in the world. This is also the case for low-level perceptual experiences such as that of pitch, heat, or pixel color. At least in our typical mathematical models, the perception of color can vary continuously; likewise, the space of possible spectral power distributions of light is also continuous.

The first section will introduce the basics of measurement theory and discuss their application to the example of color perception. The important notion of equivalent representations and invariance under suitable transformations will be discussed. I
will distinguish two crucial types of mapping in a measurement theoretic analysis of content. An $\alpha$-mapping is the mapping in virtue of which a structure represents a domain. In the case of pixel color perception, the mapping from the infinite dimensional space of possible spectral power distributions into the three dimensional space of equivalence categories induced by a color matching experiment is an $\alpha$-mapping. In contrast, $\beta$-mappings are mappings between distinct representational systems which preserve content. The content of a measurement is just whatever remains invariant under permissible $\beta$-mappings.

The distinction between $\alpha$-mappings and $\beta$-mappings will allow us to state more precisely some of the open questions in the science of color experience. For example, in Section 3.2.3, we will revisit the challenges mentioned at the end of the preceding chapter. Here, however, we will be able to state precisely the role that an answer to these questions will serve in furthering our understanding of the informational properties of color experience. A crucial distinction here will be between representational structure, structure with representational content, and artifactual structure, structure which is characteristic of the representing system but which does no informational work.

Can perception be characterized as measurement without circularity? The worry here is that measurement itself requires intentionality. If this is correct, the appeal to measurement as a foundation for a naturalistic theory of content fails as it smuggles in intentionality at the start. This worry can be avoided, however, if there is a coherent notion of natural measurement. Van Fraassen has recently argued that natural measurement is not possible. Section 3.3 refutes van Fraassen’s objection, and then turns it on its head. I will argue that in fact the very features which van Fraassen emphasizes as impediments to a theory of natural measurement can be used in the analysis of natural representational systems such as color perception in order to determine what they represent.

Finally, the chapter concludes with our first discussion of error. Explaining the possibility of error in a representational system is one of the biggest challenges for a naturalistic theory of content. Our strategy is to introduce it in stages. Here, I will introduce the distinction between precision and accuracy and argue that “error” in
the sense of imprecision can be found in natural measurements. Although inaccuracy simpliciter may not be defined within a naturalistic context, the related notion of relative inaccuracy—discrepancies between measurements of the same quantity in different contexts—plays effectively the same role and is easily defined naturalistically for a perceptual system.

3.2 Measurements and Mappings

3.2.1 What is Measurement?

Measurement occurs when a specified procedure is used to systematically relate two domains. Usually, one domain is a set of objects ordered by an attribute in nature which can vary continuously, e.g. length, temperature, hardness, hue, etc. The second domain is usually a mathematical structure, often the real line. We can demonstrate that the second domain represents the attribute which orders the first if we can show that our procedure ensures that this ordering is preserved amongst the elements mapped to in the second domain.

For example, consider the procedure of holding a thermometer up against an object and noting the height of the column of mercury. This procedure allows me to order objects: \( a \preceq b \) if and only if the height of the column of mercury is at least as high when I hold the thermometer up against object \( b \) as when I hold it up against object \( a \). If \( D \) is the domain of objects, then \( \langle D, \preceq \rangle \) is a relational structure which captures the essential empirical properties of my measurement procedure.

The formal theory of measurement examines the relationship between relational structures such as \( \langle D, \preceq \rangle \) and numerical structures such as the real line, i.e. \( \langle \mathbb{R}, \leq \rangle \). If we can produce a homomorphism \( \phi \) from \( \langle D, \preceq \rangle \) into the real line, then we will have justified our use of numbers as representations of temperatures. A homomorphism is just a structure preserving mapping. In this case, the relevant structure is the ordering \( \preceq \). For all \( a, b \in D \), we want to ensure that \( \phi(a) \leq \phi(b) \) if and only if \( a \preceq b \). The ordering relation between objects induced by the measurement procedure is then preserved in the standard less-than relation between the numbers assigned to them,
and $\phi$ is indeed a homomorphism (Krantz et al., 1971, 8–9). In general, the structure preserved by a homomorphism will depend upon the details of the measurement procedure.

The general features of measurement are not restricted to representation by numerical structures. More generally, algebraic or geometric structures may be used as appropriate. The essential feature of the measurement theoretic analysis is that the represented and the representing spaces be related by a homomorphism from the first into the second. Particularly for attributes which are naturally defined by a vector of values (rather than a single value), it seems quite natural to represent these in geometrical spaces. Consider, for example, the measurement of forces; not only do forces have a magnitude, they also have a direction and this inspires a graphical representation of forces as vectors in three dimensional space.

Some commentators do not consider geometrical representations of this sort to be measurements in the strict sense. For example, Diez (1997) argues that geometrical structures of the sort treated in Suppes et al. (1989) are not strictly speaking part of the theory of measurement. However, Diez acknowledges that these structures constitute representations, and may be grouped with measurements under a more general theory of representation along the lines developed in Mundy (1986). For the present purposes, it will be helpful to describe these relationships as measurements for two reasons. First, it reminds us that one space represents another in virtue of a mapping relationship such as that described for temperature. Second, measurement as casually understood is an intuitively plausible analysis of the role of perception. The insight presented here is not that perceptions can roughly be thought of as something like measurements, but rather that the formal analysis of measurement provides a precise way to analyze the representational content of perceptual states. So, I take it that the status of percepts as measurements is uncontroversial, and the main claim to be defended here is that homomorphisms of the kind used in measurement theory are adequate to analyze their representational content.

In fact, once one takes this perspective, it is easy to see psychophysics as the endeavor of investigating the properties of perceptual measurement scales. The methodology of psychophysics involves presenting stimuli which vary with respect to a single
attribute (the attribute to be measured by the perceptual system) and then measuring how responses to these stimuli differ. In effect, the psychophysicist assumes that experience measures the attribute presented in the stimuli, then himself measures the response of the subject in order to determine the properties of the subject’s measurement procedure.

This is a kind of reverse engineering of a measurement device. Consider, for example, the case of an ignorant scientist, who is familiar with experimental procedure, but happens to have no idea what thermometers measure. If we hand him a thermometer to play with, he might start by holding the thermometer up against various objects to see if it behaves differently. He may form a hypothesis about what the thermometer measures on the basis of its behavior in this random test. He can then systematically investigate that hypothesis by measuring the thermometer’s response in different situations. If the ignorant scientist hypothesizes that the thermometer is measuring heat, then he might hold it up against stimuli which vary only with respect to their heat and measure (say, with a ruler) the height of the column of mercury. If he performs this test, not only can he conclude that the thermometer measures heat, he can also discover features of the way it measures heat. For example, the height of the column of mercury varies linearly with respect to the change in temperature; its response breaks down in extremes of heat and cold, so there are thresholds past which this measurement procedure fails; etc.

These discoveries are similar to those which have been made in psychophysics. For example, Fechner’s Law can be seen as an hypothesis about a general feature shared by all perceptual measurements. Steven’s Power Law is the more modern hypothesis, namely that

\[ S = K I^n \]

where \( S \) signifies the strength of subjective sensation, \( I \) indicates the intensity of the stimulus, and \( K \) and \( n \) are constants which vary with the sensation being investigated. Steven’s Power Law claims that all sensation is exponentially related to the domain it measures.
CHAPTER 3. PERCEPTION AS NATURAL MEASUREMENT

For example, perceptions of brightness and loudness have an \( n < 1 \); this characterizes the fact that greater differences in stimulus intensity are required for sensation as the overall intensity increases. The difference in brightness between two lights required for me to distinguish which is brighter is much greater if both lights are bright than if both lights are dim. As discussed in the previous chapter, less than 60 photons are required to generate a perceptual experience in total darkness. In bright sunlight, far more photons than this are required to register a difference in brightness.

There are also perceptual experiences which vary in the opposite direction, namely, as overall intensity increases, less difference is required to distinguish differences in intensity. In this case, \( n > 1 \). One example here is our experience of electric shocks. The difference in pain between equal increment changes in stimulus intensity is much greater for large electric shocks than for small ones. This is perhaps unsurprising given the role of pain as a warning mechanism—greater sensation here might correspond to increasing urgency in the message being sent.

Finally, just as with thermometers (and for essentially the same reasons) there are threshold effects. At very low stimulus intensity, the stimulus simply cannot be detected at all. Our perceptual apparatus is simply too coarse to detect very fine-grained variations in our environment. At the high end of intensity, just as mercury thermometers melt at very high temperatures, our perceptual apparatus literally breaks down. Very loud noises can permanently damage your ears; very bright lights can permanently damage your eyes. It is unsurprising, then, that we cannot accurately detect these intense physical stimuli and represent them in experience.

All these considerations apply also to color. In Chapter 2 we saw that a natural way to represent lights is in terms of their spectral power distribution. Since a spectral power distribution assigns a real value to each wavelength, strictly speaking, it is a vector in an infinite dimensional space. Measurements of this space may take many different forms, some of which preserve more structure than others. The claim here is that color experience measures this space, in virtue of the homomorphism from the infinite dimensional space of possible spectral power distributions into the three dimensional space of possible color percepts, as induced by the causal interaction between photons and photopigments at the retina.
However, it should be immediately obvious that relatively little of the structure in the original space is preserved by this homomorphism. Consider the case of additivity, for example. In the previous chapter, we pointed out that additivity is preserved in this dimension reduction in a very precise sense. If we shine two test lights together, then the match produced by a subject in the matching condition will be the summation of the combination of primary lights which matches each test. Is this additivity preserved in experience, however? Is the new light produced by the subject perceived as the additive combination of his perception of the two previous lights? Note that this question certainly makes sense—we experience orange as somehow a combination of red and yellow, purple as a combination of blue and red, etc. So, there is a kind of additive structure to perceptual color space. It is extremely unlikely, however, that the apparent additive properties of experience directly mirror the additive features of the color matching experiment.

This is just to repeat the point with which we concluded the previous chapter. We have demonstrated that the exact informational content of cone firing is preserved in experience. The task of characterizing the internal structure of our perceptual experience of color (especially the similarity distances between colors) has yet to be satisfactorily completed. However, the metaphor of measurement is extremely helpful here. When I measured temperatures, I used an instrument for measuring them, namely my thermometer. In order to know the significance of the values returned from the thermometer, I need to know more about the nature of the scale being employed. Discovering that the height of mercury in the thermometer varies linearly with heat is not enough to determine the properties of the numerical scale written on the side. Yet these are what I must discover in order to assess the content of the claim that the object is $70°$.

### 3.2.2 Content as Invariance under Suitable Transformations

In the case of temperature, two distinct mappings are in common use, the Celsius and Fahrenheit scales. Both numerical representations preserve the ordering $\preceq$ induced by our measurement procedure. The Celsius and Fahrenheit scales are related by an
affine transformation: if $x$ is the temperature in degrees Celsius, then $1.8x + 32$ is the temperature in degrees Fahrenheit. This transformation reflects those properties of the real line which are relevant for the representation of temperature and those which are not. The addition of 32 shifts the origin point and multiplication by 1.8 changes the unit size. This reflects the fact that the ordering of objects established by our measurement procedure does not involve a fixed reference point (such as an origin) or fixed distances (which would fix a unit length).

The real line, however, does have a fixed origin and fixed unit size. As a consequence, relationships between different temperatures which depend upon these features of the real line will be meaningless. For example, suppose the temperature today in Houston is 100°F and that in Anchorage is 50°F. Is it meaningful to say that the temperature in Houston is twice that in Anchorage? We can see that the answer is no if we convert the temperatures into Celsius. We find that the temperature in Houston is 37.8°C and that in Anchorage is 10°C. Although 50 is half of 100, 10 is not half of 37.8. So, talk of one temperature being twice that of another is meaningless.

The usual discussion of this issue is in terms of invariance (Luce et al., 1990, Chapter 22). Celsius and Fahrenheit are related by an affine transformation, but ratios of values are not invariant across affine transformations. Ratios require a fixed origin to have meaning, but the choice of origin in our procedure for measuring temperature was arbitrary. If we want to make comparisons which will be invariant across a change in scale, we must consider ratios of intervals. Suppose in the evening Houston cools to 95°F while Anchorage cools to 40°F. Since $\frac{50-40}{100-95} = 2$, is the claim that Anchorage has cooled down twice as much as Houston meaningful? This relationship, because it is a ratio between intervals, is invariant across affine transformations. Converting all temperatures to Celsius, we see that $\frac{10-4.4}{37.8-35} = 2$. So, the claim that Anchorage has cooled down twice as much as Houston is meaningful.

The idea that content or meaning should, in general, be identified with invariance under permissible transformations has been defended at length by both Patrick Suppes (2002, and many previous works) and Robert Nozick (2001).

To understand something, we want to know the transformations it is invariant under and also the transformations it is variant under. (Nozick,
The ordinary or common meaning of invariance gives a very nice sense, qualitatively, of its technical meaning. Something is invariant if it is unalterable, unchanging, or constant. Of course, the first question that then arises is this: “What is unalterable or unchanging?” The intuitive answer, which again matches well, in a general way, the technical answers I shall examine later, is that a property of an object, collection of objects or more generally, some phenomenon, is constant. Familiar examples are the shape of a cup as we move the cup around, rotate it, turn it upside down, etc. The shape of the cup does not change, nor does its weight. (Suppes, 2002, 97)

So, we are measuring a property in the world, and the outcome of our measurement is some number on a scale. In order to understand the property represented by this number, however, we need to know the permissible transformations of the scale. Those features of the measurement which are invariant under these transformations are its content, and they correspond to features of the property in the world.

But how do we know which transformations are permissible? In the case of scientific measurement, we can derive the permissible transformations from features of our measurement procedure. In the temperature example, we formalized our measurement procedure with the structure \( \langle D, \preceq \rangle \). This is an empirical relational structure: it is a set with a relation defined on it which has been determined by an empirical procedure. The features of the empirical relational structure can be used to check the meaningfulness of features of the representing numerical structure.

Throughout these three volumes we have emphasized the concept of empirical relational structure. Objects under study are abstracted as a set \( A \), and various conclusions concerning them are abstracted as relations \( S_1, \ldots, S_n \) on \( A \). Measurement scales are merely conventional classes of representations of such an empirical relational structure, using an isomorphic numerical structure in which the objects usually are vectors in \( \mathbb{R}^m \). In such a formulation one surely must treat the defining relations \( S_1, \ldots, S_n \) as being themselves meaningful, and it seems natural to hold that any other relation that can be expressed in terms of the structure \( \langle A, S_1, \ldots, S_n \rangle \) is also meaningful. An assertion is meaningless if the attempt to express it in terms of the empirical relational structure shows it to be ambiguous. (Luce et al., 1990, 268)
In the investigation of perception, however, the situation is quite different. Consider again the example of the ignorant scientist. He does not antecedently know the appropriate empirical relational structure generated by holding a thermometer against objects because he does not know what the thermometer measures. In his attempts to discover the property in the world which the thermometer measures, he can use invariances in its behavior as evidence for the correct relational structure. In order to make these point more clearly, it will be helpful to introduce a new distinction.

We have been discussing transformations, but how are transformations characterized formally? The answer is again by homomorphisms. The affine transformation from Celsius to Fahrenheit, for example, is an instance of an isomorphism (or one to one homomorphism) because it preserves the relevant structure of the Celsius scale (in particular, ordering and relative (but not absolute) distance relationships). However, this type of mapping plays a qualitatively different role in our discussion of measurement than the homomorphisms discussed in the previous section. The latter were mappings from a represented domain into a representing domain. It is in virtue of such a mapping that the representing domain represents. Call these mappings α-mappings.

**α-mapping** – a homomorphism from one domain into another *in virtue of which* the second domain represents the first.

The mappings we have discussed so far in this section, however, are mappings between domains which represent the same thing. Only those features of a representing domain which are meaningful remain invariant across these mappings. Call these mappings β-mappings.

**β-mapping** – a homomorphism from one domain into another across which meaningful structure is preserved.

Notice that the concept of a β-mapping only applies to two domains which represent the same thing. The α-mapping determines which features of a structure are meaningful. Those features of a representing structure which are meaningful represent the features of the represented domain which determine them, this is their content. If
two structures have the same content, then there will exist some $\beta$-mapping between them across which meaningful structure remains invariant.

The example of temperature measurement is illustrated schematically in Figure 3.1. It is in virtue of the $\alpha$-mapping from temperatures into the Celsius scale induced by our measurement procedure that numbers stated in degrees Celsius represent temperature. Meaningful features of temperatures are preserved across the $\beta$-mapping from Celsius into Fahrenheit; this is because Celsius and Fahrenheit scales have precisely the same content. If we know an $\alpha$-mapping, then we can use it to derive the set of permissible $\beta$-mappings. If we don’t already know the $\alpha$-mapping, however, then we could determine some of the features of the represented structure by examining what is preserved across $\beta$-mappings.

### 3.2.3 Representational versus Artifactual Structure

When we provide a homomorphism from an ordering of objects (such as that induced by our thermometric practice) into the real line, then it constitutes a measurement if the structure of that ordering is preserved. So, those aspects of the real line which preserve structure from the measured domain are meaningful. However, since the real line is defined independently of the measurement procedure, it has additional structure. This structure is an artifact of the mathematical definition of the real line. For example, features which are particular to the origin point of the real line are artifactual in the Celsius scale. As a consequence, relationships defined between temperatures which depend in some way on this fixed origin are meaningless—they
are consequences of the real line’s antecedent structure, but not of the measurement procedure. Ratios between temperatures are an example of just such a meaningless artifact in the Celsius scale.

If domain $D_2$ measures domain $D_1$, then there is an $\alpha$-mapping from $D_1$ into $D_2$. This homomorphism will preserve some structure in $D_1$ by mapping it onto structure in $D_2$ with equivalent properties. This is the representational structure in $D_2$, and it will be preserved across $\beta$-mappings. However, $D_2$ may have additional structure, structure not referenced in the $\alpha$-mapping from $D_1$. This additional structure is the artifactual structure.

Now, if we know the $\alpha$-mapping and if we know the complete structure of $D_2$, it is easy to precisely distinguish the representational structure from the artifactual structure of $D_2$. In the case of perceptual experience, however, we do not antecedently know the complete structure of the perceptual domain, nor do we know the relevant $\alpha$-mapping.

In some cases, determining these features is relatively straightforward. Consider, for example, perception of pitch in isolation. The structure of the domain of isolated pitch percepts seems almost obvious: we naturally order pitches from lower to higher, and we naturally detect certain intervals such as the octave. This linear scale of possible pitch percepts corresponds in a natural way to vibrations in air as ordered by increasing frequency. In the case of research into pitch perception, the widespread agreement on what is being represented allowed the use of fine grained psychophysical experiments to investigate properties of the physiological response, for example to decide between the place and rate theories of neural coding in the cochlea.\footnote{Cilia (tiny hairs) in the cochlea are disturbed by vibrations in the air and trigger neurons which code the audio signal and send it to the brain. The structure of the cochlea is such that cilia in different positions are disturbed by vibrations at different frequencies. However, it is also the case that cilia can be set into vibration at different rates and this affects the rate at which the corresponding neuron fires. The place theory posits that the location of the disturbed cilia determines perceived pitch. The rate theory posits that the rate of neural firing (determined by frequency of cilia vibration) determines perceived pitch. Psychophysical evidence has demonstrated that the place theory alone is not adequate to explain the properties of pitch perception, some combination of the place and rate information is clearly being used, yet the exact mechanism is still not fully determined. For a discussion of these issues, see e.g. Chapter 6 of Moore (2004).}
but that the starting point for refining our model of pitch perception was relatively straightforward.

Arguably, the case is quite different for color perception. For one thing, there are still many challenges for producing a full color appearance model. Such a model would characterize the features of the domain of possible color percepts. Although there are many distinct candidates here, some features are broadly agreed upon. For example, the topological structure of hue space is circular. This circularity is represented by the color wheel and can be found in the roughly spherical, cylindrical, or spindle shapes of many color appearance models. It is motivated by two qualitative features of our color experience. The first is judgments of nearness and color mixture, this motivates the use of color wheels in art applications independent of their perceptual accuracy. The second is opponent color phenomena. These phenomena motivate the idea the blue and yellow are “opposites,” as are red and green. For example, no color is perceived as “in between” red and green (in the same way orange, for example, is perceived as between red and yellow). Furthermore, staring at a red patch produces a green afterimage (and vice versa). Correspondingly, blue and yellow, and red and green, fall opposite each other on the roughly circular axis of hue variation in typical color appearance models (see, for example, their locations on the CIE 1931 color space in Figure 2.7).

Is the circular structure of hue space artifactual? The general consensus, it’s safe to say, is yes. Opponent color phenomena are usually taken to be an artifact of neural wiring. They are not taken to represent any natural structure in the world independent of our perceptual system. The circularity of the hue circle itself seems obviously artifactual given the natural linear organization of wavelengths of light from lesser to greater. However, there are some wrinkles here.

First, the perceptual ordering of hues around the color circle does indeed correspond to the physical ordering of single wavelength lights in the region 400 – 700 nm of the electromagnetic spectrum (setting aside, of course, the region corresponding to purple). This can also be seen in the match between the order of colors in a rainbow and around the color circle. So, some structure in the space of possible lights is perceptually preserved in hue perception. The complications emerge when
we consider lights composed of a variety of different wavelengths (the usual case). In this situation, groups of physically distinct lights will be assigned the same perceptual experience. Two lights which fall within the same such grouping are called metamers. It appears that metameric groups cannot be defined except by reference to the features of the human perceptual system. Therefore, these groupings are taken to be completely artifactual.

Another consideration emerges from an issue which I have suppressed so far this chapter. The notion of representational content developed here is only adequate as a notion of informational content. We can examine the causal relationship between light and cones, see that it induces a mapping from lights into perceptual color experience, and interpret the content of this mapping by applying the theory of measurement. Since we are appealing to a causal process here in order to determine what is being measured, we could equally well look at earlier stages in the causal chain which ends with transduction in the retina. Most importantly, we could look at surface properties and investigate the extent to which these are being measured in experience.

The essential point here is this: if perceptual experiences are natural measurements, then any single percept may actually be measuring more than one domain. Correspondingly, some structure of the perceptual space within which that percept falls may be artifactual with respect to one domain but representational with respect to another domain. In the case of metamers, for example, the color realism debate usually treats metameric surfaces as posing the same problem as metameric lights. But this is not immediately obvious. Since the signal which impinges on the retina is produced by a combination of the spectral power distribution of the illuminant and the spectral reflectance properties of the surface, two surfaces can only be metameric with respect to a given illuminant. In fact, for any two surfaces metameric with respect to illuminant I, there will always exist an illuminant I’ such that under I’ the surfaces are perceptually distinct. In fact, it is very difficult to find surfaces which will remain metameric under many illuminants. The usual case is that relatively slight changes in illumination will reveal them as distinct.

Furthermore, we need not stop the causal chain at the spectral reflectance properties of a surface. These properties are themselves caused by the chemical structure
of the object at its surface. A metameric grouping associated with the percept green may seem unnatural if we just look at the surface reflectance properties grouped together. If, instead, we look at a more basic chemical property, the presence of chlorophyll, for example, the grouping may seem much more natural. These issues will be investigated in more detail in the final chapter. Before getting too enthusiastic about the advantages of interpreting percepts as natural measurements, it may be helpful to slow down and look at some potential objections to natural measurement as a coherent theoretical concept.

3.3 Distortion in Natural Measurement

3.3.1 Natural Measurement

In the case of scientific measurement, a scientist performs a measurement procedure and this procedure induces a homomorphism from an empirical domain into a mathematical structure. In the interpretation of perception discussed above, a causal process induces a systematic relationship between an external domain and an internal domain, in particular a level in the hierarchy of perceptual experience. If this analysis works for perception, however, there’s no reason it shouldn’t work for other domains which are systematically related by a causal relationship. This is natural measurement.

To be somewhat more precise, call a set of mutually exclusive, extensionally defined, and continuously varying properties a quality domain. Under this definition, length, weight, position, temperature, speed, degree of reflectance, rate of fire, and many other features of the natural world constitute quality domains. Lengths, for example, are mutually exclusive: if an object is length \( x \), then it cannot also be length \( y \neq x \). The mutual exclusivity of lengths is a consequence of their extensional definition; lengths can be identified with sets of objects, the ends of which line up when they are held next to each other. Finally, lengths vary continuously: if \( x \) and \( y \) are lengths such that \( x < y \), then there is a length \( z \) such that \( x \leq z \leq y \). \(^2\)

\(^2\)The basic features of natural measurement as discussed here could be reconstructed without the
We need one more concept before we can define natural measurement, which I will call *environmental constraints*. Most of the interesting examples of natural measurement do not hold in an absolute sense, but only relative to certain environmental conditions. If these conditions obtain, however, they constrain the relationship between two quality domains such that information can flow from one to the other. In the words of Barwise and Perry (1999): “These constraints are what provide reality with a structure that supports the flow of information” (97). For example, the orientation of a column of smoke varies systematically with the direction of the wind and thereby constitutes a natural measurement of wind direction. However, this measurement relation can be defeated if the causal connection between wind direction and smoke column orientation is broken, say by placing a battery of high powered fans next to the source of the smoke.

Barwise and Perry call these *conditional constraints*, and emphasize that the crucial point about such constraints is not that they be explicitly recognized, but that they obtain. If I am a tracker who wants to obtain information about wind direction from observing a column of smoke, my success or failure in this endeavor depends upon whether the relevant constraint obtains, *not* on my awareness of the constraint or how it might be defeated.

An important point to notice here is that as long as we stay within the appropriate environment, we needn’t be aware of the conditions at all. Birds don’t need any attunement to glass as long as they stay in an environment where there is no glass. . . . It is only when one wanders out of one’s setting that it becomes important to be attuned to the conditions under which the old constraint remains applicable. (Barwise and Perry, 1999, 100)

Barwise and Perry say agents are “attuned to a constraint” if they acquire information from relations which hold conditionally with respect to the constraints. This basic features of information using organisms has been recognized and discussed by others as well. Millikan, for example, discusses essentially the same point in terms of “tracking”
a (potentially) “gerrymandered geographic domain” within which an informational relationship holds. However, the discussion in Chapter 5 of Barwise and Perry (1999) is the most simpatico for the present purposes.

A crucial point here is that constraints and the informational relationships which hold relative to them can be read off of a description of the world. They should not be confused with an attempt to define the normal or standard conditions, which are famously problematic for naturalistic theories. We can immediately see this from another simple example. The length of a shadow measures the height of an object relative to the constraints that (i) there be a single illuminant (say, the sun), and (ii) the position of the illuminant is fixed (in the case of the sun, its apparent position). But no (apparent) position of the sun in the sky (or of any other illuminant, for that matter) is more typical or normal than any other. All that matters is that various positions do obtain and, given that some particular position obtains, lengths of shadows measure heights of objects.

We are now in a position to define natural measurement:

natural measurement – a quality domain \(X\) naturally measures the quality domain \(Y\) relevant to constraints \(C\) if, whenever \(C\) obtain, the causal relationship between \(X\) and \(Y\) induces a homomorphism from \(Y\) into \(X\).

The standard statistical treatment of measurements is in terms of random variables. Remember from the discussion in Chapter 1 that a random variable is just a function on a sample space. Natural measurements can also be statistically modeled with random variables: the represented domain \(Y\) is just the domain of the random variable, or the sample space; the representing domain \(X\) is just the range of the random variable. As we will see in our discussion of error, no assumptions need be made about the procedure which defines this random variable in order to analyze properties of the data it generates.

### 3.3.2 The Objection from van Fraassen

The point of interpreting perceptual experience as a natural measurement is to provide a naturalistic analysis of the representational features of perceptual experience.
If the attempt to interpret perceptual experience as a measurement smuggles in intentionality at the start, then it does not succeed as a foundation for a naturalistic theory of intentional content.

In his most recent book, van Fraassen explicitly discusses the role of measurement as a means of representing phenomena. Van Fraassen argues that measurement is inherently perspectival and theory-laden. These features are derived from its inherent intentionality. Van Fraassen’s argument depends upon drawing an analogy between features of representation in art and features of representation in measurement. In this section, I outline van Fraassen’s argument and demonstrate some crucial dis-analogies between the perspectival features of artistic representations and those of scientific measurements. I next demonstrate that the perspectival features of scientific measurement have analogs in natural measurements which are subsumed under the environmental constraints.

Van Fraassen (2008) begins with an analysis of representation, concluding with a “fundamental theorem” of representation:

There is no representation except in the sense that some things are used, made, or taken, to represent some things as thus or so. (23)

A consequence of this fundamental theorem is that it “leaves no room for ‘representation in nature’, in the sense of ‘naturally produced’ representations that have nothing to do with conscious or cognitive activity or communication” (24). Van Fraassen next argues that measurement is a form of representation. Since there can be no natural representation and measurement is a species of representation there can be no natural measurement.

Obviously, I agree that measurement constitutes a form of representation. However, I think it is wrong to move from an analysis of representation in a general sense to claims which supposedly apply to any particular theory of representation. In the case of measurement as discussed above, measurement as a relationship between two domains is defined first. It then follows from the nature of this relation, namely that the structure of the first domain is re-presented in the second, that the measuring domain represents the measured domain.
The weak link in van Fraassen’s argument is the extension of his fundamental theorem to all forms of representation. The considerations which motivate the fundamental theorem concern the nature of representation in art. Here his approach closely follows that Goodman (1976). Goodman begins by emphasizing that resemblance is not enough for, nor even essential to, representation. Two paintings may resemble each other quite closely, in the sense of exhibiting many of the same physical features in the same arrangement, yet represent quite different objects. Likewise, even though one painting may resemble an object more closely than another one does, this other painting may nevertheless constitute a better representation of that object.

Goodman argues that representation can be understood better by examining the notion of representation-as. Pictures do not represent simpliciter, they represent an object as being thus-and-such, e.g. Bush as class dunce, or Obama as the Joker. Van Fraassen extends Goodman’s discussion of this topic by emphasizing that distortions (deviations from strict resemblance) can serve to make representations more successful, as in the case of caricatures. In this case, misrepresentation is also a species of representation. A representation of Obama as the Joker may be successful in the sense that viewers clearly identify the subject as Obama and his characteristics as those of the Joker. But it may also be a misrepresentation in the sense that it falsely portrays Obama as having Joker-like characteristics (van Fraassen, 2008, 13–5).

The moral here is that representations are inherently perspectival or indexical. There is no representation simpliciter, but only representation from a particular perspective.

Let us grant van Fraassen this point. Nevertheless, there is an important difference between the perspectival features of a caricature and those of a scientific representation. Let’s compare two concrete examples: a caricature of Nixon and a Mercator projection.

The caricature of Nixon is distorted in a fashion which emphasizes the most salient features of Nixon’s face. His nose, for example, is noticeably pointy and his jowls noticeably slack. The clever caricaturist takes advantage of these salient features, extending the nose to an absurd length and emphasizing the slackness of the jowls. Conversely, the rest of the head may be de-emphasized, even shrunk. This artistic
distortion of Nixon’s features makes his depiction quickly recognizable to anyone familiar with Nixon’s face. It depends both on a common tendency in human attention to facial features, and on particular choices on the part of the artist about just how much to emphasize or downplay this or that feature.

A Mercator projection is a two dimensional map of the features of the globe. Take a transparent globe with key features marked upon its surface in black (the outlines of continents, locations of cities, etc.). Now, wrap a sheet of paper around the equator. Notice that the paper and globe will touch only at the equator—the distance of the globe’s surface from the paper increases to the north and south. Now, shine a light from inside the globe and trace the shadows cast by its markings onto the surface of the paper. This is a Mercator projection.

A Mercator projection presents a distorted depiction of the surface of the globe. Furthermore, the degree of distortion changes as one moves further from the equator. Eventually, at the poles, the distortion becomes infinite, and a single point on the globe (the north pole, say) is represented as a whole line on the map. In fact, the top and bottom edges of the Mercator projection will correspond to the north and south poles respectively.

The Mercator projection is also inherently perspectival. We chose a particular perspective by choosing to wrap the paper around the equator. We could have chosen any other circumference, however, and performed the same procedure. The choice of which circumference around which to wrap the paper determines which parts of the map will be more distorted and which less distorted. It is human purpose, the intent of the mapmaker, which dictates the equator as the most salient placement of the paper. This placement maximizes the utility of the map for navigational purposes.

However, there is a key difference between the distortions of the Mercator projection and those of the caricature of Nixon. Once the perspective is set, the distortion of the Mercator projection proceeds mechanically and automatically. The degree to which a particular segment of the globe will be distorted in the projection is not a matter for choice—once the initial decision of which circumference along which to wrap the paper is made, all the remaining details of the mapmaking process are completely determined.
The caricature of Nixon is totally different in this respect. After the artist chooses which features to emphasize, he still has many remaining degrees of freedom. All caricatures of Nixon emphasize the length of his nose, yet the degree to which the nose is exaggerated differs with the artist. Furthermore, the emphasis on Nixon’s jowls can be determined independently of other features of the caricature. Once the length of Nixon’s nose is set, the width of his jowls does not follow by mechanical procedure or set of rules, but remains at the whim of the artist.

So, while acknowledging that scientific representations are inherently perspectival, we can nevertheless maintain that this indexical feature is importantly different from that found in artistic representations. Perspective of the sort found in the caricature of Nixon will never be found in nature because intent and whim permeates every detail of the representation. But perspective of the sort found in the Mercator projection may very well be found in nature. Once the orientation of the projection is determined, its details follow in a mechanical, even causal manner. So, if there are instances where a perspective itself can be determined by a natural causal process, these may be qualitatively similar to that found in the Mercator case.

In fact, we have already seen such an example in the previous section. The measurement of the height of an object by the length of its shadow is inherently perspectival; it depends crucially on the position of the sun. The position of the sun indexes the measurement of heights by shadows in exactly the same fashion as the choice of where to wrap the paper around the globe indexes the Mercator projection. In both cases, once the initial position is set, the relationship between measured (globe, height) and measuring (paper, length) follows in a systematic fashion from a causal process. Furthermore, the placement of the sun itself is determined by a natural process (the rotation of the earth), not by any smuggled in intentionality or purpose.

Ultimately, the disagreement here is essentially one of terminology. Van Fraassen and I disagree about the conditions which need to be met for representation to occur. I take the re-presentation of information (with measurement relations a specific instance of this) as the defining feature. Van Fraassen takes the presence of intentionality or purpose as necessary for representation to occur. This does not mean that he denies
CHAPTER 3. PERCEPTION AS NATURAL MEASUREMENT

the existence of the types of natural relationships which I have defined here as natural measurements:

On the macroscopic level too we can think of processes that connect two situations separated in time or space. These could be so correlated that it would be possible in principle to get information about the one by inspecting the other—provided of course we knew of that correlation! But that something could be done does not mean that it is done. That something could be assigned a representation role does not mean that it has one. There are contexts where the distinction between a measurement and its physical correlate can be neglected, but philosophical reflection is not one. (van Fraassen, 2008, 156)

My view is that van Fraassen simply begs the question against the naturalist here. Our project is to define and explain the characteristic features of intentionality using only descriptive features of the world. It constitutes a step forward in this project to realize (i) that relations which share qualitative features with scientific measurements can be found in nature, and (ii) correspondingly, the definition of representation used in measurement theory applies mutatis mutandis to these natural relations. In this case, philosophical reflection shows us that the physical correlates of measurement are themselves measurements in a very precise and specific way.

3.3.3 Turning van Fraassen on his Head: Distortions as Evidence

Perhaps the disagreement between van Fraassen and myself is merely one of terminology (or agenda). Nevertheless, there are additional issues in van Fraassen’s discussion it may be helpful to address here. The first is the inherent “theory-laden” character of measurement; the second is the value of distortion, or departures from strict resemblance, in effective measurement. I hope to dismiss the relevance of the first issue to the present project rather quickly (acknowledging that a deeper discussion is warranted in the context of philosophy of science). The second issue I hope to turn on its head, demonstrating that it is not a worry for the perspective introduced here, but rather a valuable source of evidence.
Van Fraassen’s discussion of the inherent theory-ladenness of measurement plays a deep role in his defense of his current position in philosophy of science, empiricist structuralism. The aim of the present work is not to develop any general theory about the nature of scientific knowledge, so the specifics of van Fraassen’s view need not concern us here. It will be helpful to show, however, that van Fraassen’s worry can be seen as the converse of one of the worries we leveled against the teleological approach in Chapter 1.

One example van Fraassen considers is the historical development of the measurement of temperature. His agenda is to defend the claim that our mature theory (and, correspondingly, measurements) of temperature can not be understood as latching on to some theory-independent feature of the world. In order to defend this claim, he emphasizes the role at each stage in the evolution of thermometry of convention, choice, and the intention of the researcher. “[A] definite identification, a complete definition, of the measured parameter is possible but only through, at the hands of, and relative to the theory offered and finally accepted to account for the stability of the measurement procedure.” So, the success of our contemporary practice of thermometry depends upon two types of facts:

(a) an empirical fact that has been discovered: namely, the very stability in the procedures found in this historical development, and the reliability of the predictions concerning these and their correlation with other measurement procedures derived from the mature theory in which they are now theoretically embedded.

(b) a historical fact: namely, that choices in the development of these measuring procedures went hand in hand with the development of the theory in question, so that we cannot identify an aspect of nature that is measured if we refer only to those empirical procedures without using concepts provided by the theory. What now counts as simple passive measurement is a hard-won achievement. (van Fraassen, 2008, 124)

The empirical facts to which van Fraassen refers are of course the types of stable causal relations between domains I have called natural measurements. Now, does an analysis of the choices made during the development of thermometry demonstrate
that contemporary measurements of temperature are inherently theory laden? I wish to remain agnostic on this point in the present work. In general, I am sympathetic to many features of the position van Fraassen develops. However, a far more nuanced and sophisticated analysis of the history of thermometry has been given by Hasok Chang. At least one feature of van Fraassen’s analysis is refuted by the discussion in Chang (2004): namely, the claim that the improvement of a measurement procedure requires a theory of what is being measured. Chang’s discussion of Regnault’s experiments on thermometric measurements in Chapter 2 conclusively refutes this claim in my opinion.

Nevertheless, there is an essential point here to which I am deeply sympathetic. van Fraassen argues that since intentionality (in the form of theoretical bias) featured in every stage of the historical development of thermometric measurement, our current practice of thermometry is infected with this intentionality. In Chapter 1, I argued that the converse is true of teleological arguments: if the historical development of a structure is the source of its intentionality, then there must have been intentionality at each stage along that development. I was worried that the appeal to evolutionary pressures in the definition of representational function depends upon an illegitimate attribution of intention to evolutionary forces themselves. At the very least, van Fraassen and I can agree on this: once intentionality enters the picture, it infects every endeavor which it guides. The naturalistic challenge is to define and explain the emergence of intentionality in natural systems without reference to prior intentional structures.

Another crucial feature of the intentionality of representations for van Frassen is the role of distortions. We saw in the previous section that distortions can be the consequence of perspective. This is true even in a very natural and intuitive sense: my perspective on the train tracks distorts their properties such that they appear to converge. From my perspective, this table appears to be a rhomboid although I know its corners form right angles. One feature of distortions which van Fraassen emphasizes is that distorted representations can be better representations.

Van Fraassen begins his discussion of distortion with an anecdote from Gombrich concerning a competition between two ancient Greek sculptors. The competition is
to produce a sculpture of Minerva to place upon a high pillar. Alcamenes produced a sculpture of a beautiful woman, while Phidias produced a distorted sculpture which stretched and enlarged the features of the face. Yet, when the sculptures were placed upon the high pedestal, that of Phidias was much more pleasing to the eye when viewed from below (Gombrich, 1960, 191). Van Fraassen then argues that similar distortions may be found throughout art, concluding:

> It seems then that distortion, infidelity, lack of resemblance in some respect, may in general be crucial to the success of a representation. This does not rule out that resemblance in some other respect may be required. Yet even when that is the case—and it may be a special case—the choice of those respects in which resemblance or a specific kind of distortion is required, and those for which just anything at all will do, will have to be seen as crucial as well. (van Fraassen, 2008, 13)

Now, in the case of natural measurements, we have no resources to evaluate them as good or bad, effective or ineffective on the basis of representational function. However, we can identify distortions. I propose that van Fraassen’s insight here—that distortions can improve the quality of representations—can be used to help us solve the problem of deriving a naturalistic analysis of intentional content. If we identify a distortion in the representational properties of a perceptual space, we can use this distortion as evidence for determining what the space represents.

Consider the example of electric shocks discussed above. The intensity of the experience of a shock increases exponentially with the increase in the electrical current. We can determine this empirically. Furthermore, the experience itself is negative, again, a simple empirical determination. Now, the response to an electric signal is a distortion in van Fraassen’s sense: because the scale is exponential, smaller changes in current are represented as bigger changes in experience as the absolute value of the current increases. So, as a measurement of the absolute value of electrical current, our perceptual experience is rather poor. But, as a measure of the danger to the body, and the urgency of removing it from the electrical current, the scale is quite effective.

In this case, it seems plausible to derive from features of the perceptual experience of electric shock (the experience is negative + it increases exponentially with the
strength of the shock) its function to warn us of danger. Now, this explanation is still very schematic. But, if we could show in a precise way that the experience we feel from shock contains more information about danger to the body than about the strength of the electrical current, then this would give us a basis for determining the representational function of this experience from a purely descriptive account of its informational relationship to stimuli.

Finally, let’s return again to our central example, color. The space of color experience is a distortion of the space of possible wavelengths of light. We know that cone cell response measures wavelength of light because we understand the causal process by which this occurs. We don’t know the causal process which leads from cone firing to the experience of color, but if we assume it, we discover (as demonstrated in Chapter 2) that the information about wavelength which can be found in cone firing is exactly equivalent to that recoverable in a color matching experiment.

However, once we discover that experience of color measures wavelength, we can examine the similarities and differences between the measured and the measuring domains. We’ve already mentioned the problematic circularity of hue space. Another crucial feature here is the distorted representation of the spectrum in color experience. We know that the response functions of cone cells are not evenly spaced across the visible spectrum (see Figure 2.4). Cone cells recover more information, then, in the range of 550 – 650 nm than they do in the range of 450 – 550 nm. We would expect a corresponding distortion in perceptual color experience, namely more fine grained assessments of color information in the Red / Yellow / Green region of color experience than in the Blue / Purple region.

In fact, we do find something like this, as can be seen in the Munsell color space and many color appearance models. There is more detail (more distinguishable variation in possible perceptual experience) in the Red / Yellow / Green region than in the Blue / Purple region. Now, we know that color experience potentially represents anything in its past causal chain. We can looks at steps in this past causal chain and assess the value of the distortions in perceptual color space for representing information in those steps. This style of analysis might provide additional support for a thesis such as that advanced at the end of Section 3.2.3. If the distortions in perceptual color
experience can be shown to represent information about chemical composition more effectively than information about surface reflectance properties, this consideration may be used to argue in favor of the representation of chemical properties as the primary representational function of color experience.

3.4 The Problem of Error, part 1

3.4.1 Precision and Accuracy

As we saw in Chapter 1, one of the major challenges for a naturalistic account of intentional content is to explain representational error, or misrepresentation. It is a mistake, however, to treat error as a single phenomenon. There are distinct types of error and some pose more of a problem to the naturalistic approach than others. In this section, we examine two types of error, random error and systematic error, and when they emerge in natural measurements. Random error, at least, is always present; realization of this should satisfy those who take the possibility of misrepresentation as a necessary condition for representation that natural measurements indeed produce representations.

One of the advantages of examining continuous representational spaces is that we are not restricted to binary evaluations of representational success. We can say more than just: this measurement outcome is correct / incorrect. We can speak of degrees of correctness, of measurements being better or worse. Since we are examining two types of error, we need two distinct notions for evaluating the quality of a measurement. These are precision and accuracy.

- **precision** – the smaller the random error in a measurement procedure, the higher its precision
- **accuracy** – the smaller the systematic error in a measurement procedure, the higher its accuracy

It is important to understand that at least some random error is present in any measurement procedure. This is true for both scientific measurement and natural

---

3Definitions adapted from Beers (1957), 4.
measurement. This is because conditions in our world are in constant flux, and no two causal interactions ever occur in precisely the same way. Consequently, the exact cause of random error on any individual occasion is beyond human ken. However, we can analyze the random error in a measurement procedure by repeating that procedure. The values of this repeated measurement will never agree precisely. The extent to which they do agree is a measure of the precision of the measurement procedure being employed.

Beers lists four causes of random errors:

1. *Errors of judgment*. Most instruments require an estimate of the fraction of the smallest division, and the observer’s estimate may vary from time to time for a variety of reasons.

2. *Fluctuating conditions* (such as temperature, pressure, line voltage).

3. *Small disturbances*. Examples of these are mechanical vibrations or, in electrical instruments, pickup of spurious signals from nearby rotating electrical machinery or other apparatus.

4. *Definition*. Even if the measuring process were perfect, repeated measurements of the same quantity might still fail to agree because that quantity might not be precisely defined. For example, the “length” of a rectangular table is not an exact quantity. For a variety of reasons the edges are not smooth (at least if viewed under high magnification) nor are the edges accurately parallel. Thus even with a perfectly accurate device for measuring length, the value is found to vary depending upon just where on the cross section the “length” is measured. (Beers, 1957, 5)

Obviously, the first of these does not apply in the case of natural measurement: there is no agent wielding the measuring device who might make such an error in judgment. 4 is somewhat puzzling; I believe a full analysis of the relationship between this issue and natural measurement would take us too far afield, so I will simply set it aside. 2 and 3, however, are purely features of the causal structure of the measuring situation. Consider again the case of the length of a shadow as a natural measure of the height of, say, an obelisk. The edge of this shadow will be fuzzy due to diffraction, the height of the obelisk will vary with temperature due to thermal expansion, elephants lumbering past nearby may produce enough vibrations to disturb
the exact location of the shadow’s edge, etc. Even if we take the position of the sun as fixed, and even if the shadow of the obelisk is produced by it due to a purely causal process, imprecisions will enter into this measurement. Furthermore, these imprecisions could be discovered by comparing the length of the shadow as cast at the same hour of the same day from year to year. The cycle of the sun’s position in the sky is a year long, so at the same hour of the same day each year, the sun should be in the same place. Even here, however, is a source of imprecision since, strictly speaking, even our yearly journey around the sun (and correspondingly the sun’s position in the sky) is subject to subtle variation and change.

What about our experience of color? The imprecision in the physiological measure of light can easily be seen in Figure 2.2 (for rods) and Figure 2.4 (for cones). If wavelength is what is being measured here, it is being measured with extreme imprecision. We also find imprecisions in the perceptual response. In general, these imprecisions emerge as variations in the response to psychophysical tasks. In essence, the psychophysical measurement of a perceptual response inherits imprecision from imprecisions in the response itself. For example, if I perform an experiment to determine the softest tone you can hear of a specified pitch, the result will vary from trial to trial and with the experimental setup. This shows that I can play the same (very soft) stimulus to you and on some occasions you hear it and others you do not. This is just more evidence for imprecisions in the perceptual system.

If all measurements exhibit some degree of imprecision, and if all physical systems are subject to random fluctuations, can we make sense of the notion that there is a “true” value for some quantity in the world? In the statistical analysis of data, the “true” value is usually just a theoretical entity, helpful because positing it suggests useful strategies for interpreting data. If, however, we intend to measure a specific quantity, then our intent can define the properties of that quantity well enough to determine if our measurement procedure is appropriately directed toward it. The worry here is that our measurement procedure may contain some flaw which systematically biases it away from the true value.

For example, suppose I am measuring temperatures with my thermometer. I report that the water is 65°Fahrenheit, but you measure it with your thermometer
and report that it is 60°. Surprised at the discrepancy, we measure a number of different objects and discover that each time my thermometer reports temperatures to be 5° above those reported by yours. One of our thermometers has been miscalibrated, and this miscalibration produces a systematic bias in the measurements. Of course, the physical relationship between mercury and temperature has not changed—so the height of the column of mercury is just as accurate a measure of temperature as it ever was. The error is produced because the numbers have been placed on the thermometer in the wrong place, producing a systematic bias away from the “true” value as defined by the intended application of the Fahrenheit scale. So, the inaccuracy of the measurement derives from my intent to measure with the Fahrenheit scale plus the miscalibration of my thermometer with respect to that scale.

We have seen that natural measurements are subject to errors in the sense of imprecisions. This example, however, makes it appear that they cannot be subject to errors in the sense of inaccuracies. This is because the notion of accuracy depends upon a notion of the “true” or intended value a measurement procedure is directed at. Yet this is precisely what we do not have in the case of natural measurements. In the special case of perceptual experience, however, there is a strategy for naturalistically determining if experience is somehow miscalibrated. The strategy will be to compare different levels of experience which measure the same domain, and thereby determine their relative calibration. This will allow for a definition of systematic bias in the perceptual system which does not illegitimately import a prior notion of intentionality and, consequently, can be used as the first step towards a naturalistic definition of intentional content.

3.4.2 Inaccuracy in Perception

Weber’s experiments on weight perception demonstrated that our experience of weight is calculated from a number of lower levels in both the hierarchy of neural processing, and the perceptual hierarchy which corresponds to it. Let’s remind ourselves of the experiment which demonstrates that perception of temperature contributes to the perception of heaviness. Weber alternately placed a pair of silver Thaler at room
temperature and a single silver Thaler at near freezing temperature at the same location on a subject’s forehead. When the subject was asked to make a judgment as to which was heavier, he reported that the two stimuli were the same weight, or even that the single, cold Thaler was heavier. Is this an example of inaccuracy in perception?

First, if we have a basis for claiming that the experience of heaviness is about weight and nothing else, then we can recognize this as an instance of inaccuracy. But, as this same experiment itself demonstrates, our experience of heaviness contains information about not only weight, but also stages in the causal chain leading to that experience, in particular, information about temperature and pressure. We need to distinguish weight from the other attributes that the experience contains information about. A strategy to do so which has been mentioned above, but not yet developed in detail, is to demonstrate that perception of heaviness contains more (efficient) information about weight than about any other step in the causal chain which produces it. This is the strategy which will be developed in Chapter 6.

However, in this simple example, we do not need anything so sophisticated. In fact, careful examination reveals that this example is actually quite similar to the thermometer example discussed above. If I place successively more Thaler on your forehead at constant temperature, you perceive them as increasing successively in heaviness. Once we realize this, we can see that temperature acts as a kind of calibration on the measurement of weight from pressure. Perceptions of weight at lower temperatures and perceptions of weight at higher temperatures merely utilize different scales. Through careful experimentation, Weber has discovered a way to measure the difference in calibration between two specific scales amongst many here (roughly equal to the weight of a single silver Thaler), just as through experiment, we were able to discover the difference in calibration between our two thermometers (5°).

Now, we can immediately conclude that one or both of these two scales has systematic bias. What we don’t have is a grounds for saying that one scale is correct and the other incorrect. But we don’t need one! This is because the assessment that one Thaler is heavier than two Thaler is incorrect on either scale. We don’t get to conclude that one scale is inaccurate and the other accurate, but we do get to conclude
that, insofar as they measure the same thing, they differ in accuracy. Furthermore, when data is intermingled from each of the two scales, comparisons of data become inaccurate.

This same strategy can be used to explain many of the incorrect judgments which emerge in optical illusions. Consider the Müller-Lyer illusion, for example (Figure 3.2). We experience the line with outward pointing ends as longer than the line with inward pointing ends, even though they are exactly the same length. If we consider only lines with outward pointing ends, however, we have no problem ordering them correctly as to length. Likewise with lines which all have inward pointing ends. So, we can derive the error of the judgment that the two lines are equal from descriptive facts about perceptual experience. We don’t need an external standard of correctness because of the agreement between scales for each condition considered separately. Since both scales agree that the lines are equal, we can see that the appearance of inequality is an error, an inaccuracy, or a miscalibration in the perceptual scale which emerges during the perception of line stimuli with mixed inward and outward pointing ends.

In the case of both Weber’s experiment and the Müller-Lyer illusion, an apparently irrelevant factor turned out to calibrate our subjective scale for measuring an objective quantity in the world. Since the shift in calibration was fairly straightforward, we were able to recover the “correct” value from the agreement between scales at either calibration. In the case of color perception, however, the story gets somewhat more
complicated. Once we consider our perceptual experience of color in context, it is clear that the context in which a color stimulus is presented calibrates our scale for measuring it. In the simplest case, context is just the reflectance properties of the background on which the target patch is placed (see, for example, Figure 3.3).

Now, on the one hand, it is easy to see this as calibration. The lightness and hue of the background affects the appearance of a colored patch. The orange square seems muddy and brownish against a light yellow background, but brighter and more vivid against a dark green background. What we don’t get in this example is agreement across all different calibrations with respect to the measurement of some easily identifiable external property. If we hold hue fixed, we can get it for brightness. But if we allow hue, saturation, and brightness to all vary, then we will get different results in different contexts. Part of the problem here is that the complicating factor is not obviously irrelevant. In both the Müller-Lyer case and the weight judgment case, the complicating factor is clearly a distinct property, and thus easily corrected for in order to arrive at more fundamental scales. We correct for temperature and find the judgments of heaviness have the anticipated properties. We correct for arrows at the end of lines (or remove them entirely) and get a more primitive, and effective, perceptual measurement of length.

Color is different, though. Here the complicating factor is just color again. Furthermore, we’ve already seen that our experience of colors in context and that of pixel colors are qualitatively different. We cannot then, in a straightforward and satisfactory way, appeal to the appearance of the color in isolation as its “correct”
color. Instead, we will need to pursue our earlier suggestion in more detail: look for the properties in the world about which color experience delivers the most efficient information.

3.5 Conclusion

In this chapter we defined natural measurement and demonstrated that perceptual experience constitutes a natural measurement of properties in the world. We considered and refuted an argument against the possibility of natural measurement, and we demonstrated that, contra expectation, errors in natural measurement can be defined without illegitimately importing intentionality. In particular, natural variance in our complex world ensures that all natural measurements exhibit some degree of imprecision. In the special case of perceptual measurements, we can define relative inaccuracies which in some select cases can be used to define inaccuracy simpliciter, again without resort to intentionality.

In the specific case of color perception, however, this is not so easily done. Although relative inaccuracies are easy to find, there is no clear cut story of inaccuracy simpliciter for color experiences. Nevertheless, in the following chapter I will demonstrate how a naturalistic semantics for color terms and concepts can be provided which allows them to be detached from the causal chain between world and experience. This solves the most pressing challenge for a naturalistic account of intentional content by allowing for color concepts to be tokened in error when no corresponding color stimulus is present. In Chapters 5 and 6, then, we return to the issue of inaccuracy and develop an apparatus for determining that space of variance in the world represented most efficiently by color experience.
Chapter 4

Discretization and Detachability

4.1 Introduction

At the end of the previous chapter, we examined how two types of error may emerge in the natural measurements made by the perceptual system: imprecision and inaccuracy. Although these are both “error” in the sense of deviations from veridicality, they are not enough to explain the types of error that arise with full-blown intentionality.

Imprecision is a fuzziness in measurements; there is more imprecision the wider the spread of data when a measurement is repeated. Inaccuracy is a systematic distortion in measurement; we saw in the last chapter that we cannot speak of inaccuracy simpliciter, but only relative inaccuracy when examining natural measurement systems. Inaccuracy and imprecision in this sense both arise when there is a direct causal connection between the perceptual experience which measures and the external domain which is measured.

Suppose now that I say “John’s car is red,” but I have never seen John’s car. Suppose also that John’s car is a deep forest green. Then my claim is in error, but it is not an instance of either inaccuracy or imprecision. My error derives not from noise in my causal connection with John’s car, but because I have no causal connection with John’s car at all. In order to explain how such errors may occur, we need a theory of detachability: how can mental representations become detached from the properties in the world which they represent?
CHAPTER 4. DISCRETIZATION AND DETACHABILITY

It is important to understand the challenge here. I will not attempt to give a story about how a belief such as “John’s car is red” may be tokened in the mind; this is too complex a task to accomplish here. I will focus merely on providing a naturalistic semantics for color terms like “red”. The problem is this: thus far we have explained content in terms of a direct causal relationship between perceptual system and world. My experience of color represents types of light because it measures light via the causal interaction at the cone cells. But the content of the term “red” cannot be derived in this way since it can be tokened in instances like the above where no causal connection is present.

The essential strategy here is to detach semantics from aboutness. The semantics of color terms derives from their relationship with the space of possible color percepts. But this relationship does not depend upon the relevant percept being tokened in order to hold. So a term like “red” can be used to say of an object that it will trigger the perceptual experience of redness without that object in fact being present (or being red).

Section 4.2 will give some theoretical motivation for this strategy due to Ray Jackendoff. Although I only consider color terms in this chapter, my hope is that this strategy will extend in a natural way to other perceptual domains. Obviously, even if this endeavor is successful, it will not vindicate Jackendoff’s entire research program. Nevertheless, his ideas will at least serve to motivate the strategy I implement here.

Section 4.3 will outline a formal model for detachability due to Gärdenfors. This model proposes a concrete algorithm for extracting discrete representational spaces from continuous representational spaces. Since Gärdenfors has provided a concrete algorithm, I argue that his account satisfies the naturalistic desideratum that explanation must be derived from descriptive facts about the world. Of course, it is one thing to propose a process which may occur within the brain and another thing entirely to actually discover that process in the brain. Section 4.4 will examine some recent cross-linguistic empirical findings which support a Gärdenfors-type model.

Although the discussion throughout will focus on the relationship between color terms and perceptual color space, it is important to notice that nothing about the
story concerning detachability here depends upon language in particular. The recollection and categorization of colors may also depend upon some discretization of color space. Exactly how this discretization is affected by language is a complex topic which we will not investigate here. The essential point is just that if the solution to the problem of detachability presented here succeeds, then it will apply also to the detachment of non-linguistic internal representations of color space, e.g. as in imagining or dreaming about colors.

4.2 The Problem of Error, part 2: Internal Semantics as Solution to Detachability

Ray Jackendoff was one of the most significant figures in the generative syntax movement. In the 1970’s, the so-called “linguistics wars” raged between the Chomskian camp, which privileged the study of syntactic structure, and the generative semanticists, who postulated that all surface syntax forms were derived from underlying semantic structure. While Chomsky launched broad attacks against the generative semanticists, promising a new formal apparatus, Jackendoff was often the one working behind the scenes, fleshing out the details of Chomsky’s suggestions (see Harris, 1993, 170–4). As such, no one is more suitably placed than Jackendoff to defend the importance of semantics for even the Chomskian approach to linguistics.

One of the most significant features of the Chomskian movement was its founder’s insistence that the focus of linguistics be “I-language”, the internal mental structures which generate and interpret language, rather than “E-language,” that external system which arises as a social construct within a communicating community. Jackendoff extends this attitude to the study of semantics, taking as his starting point the “uncontroversial postulate” that “People find sentences (and other entities) meaningful because of something going on in their brains”. The more controversial move this platitude inspires is that “we are ultimately interested not in the question: What is meaning? but rather: What makes things meaningful to people?” (Jackendoff, 2002, 268) Jackendoff argues that semantics should investigate the internal structures and
relationships which induce meaningfulness for humans rather than, say, reference, the problem of how these internal structures relate to the external world.

One way of interpreting Jackendoff’s suggestion is not as a rejection of reference, but rather the breaking down of the problem of reference into two distinct steps. One step concerns the relationship between the external world and our internal representations of it; the second step concerns the relationship between these internal representations and combinatorial linguistic structure. Whether this strategy will work in general is an open question. Certainly, Jackendoff’s own approach to semantics has not succeeded in supplanting all other contenders. However, in the specific case of perception and perceptual terms, this strategy looks especially plausible.

The interfaces to the perceptual systems are what permit one to form a thought based on observing the world . . . In turn, by using such thoughts as the input to language production, one can talk about what one sees, hears, tastes, and feels. (Jackendoff, 2002, 273)

Jackendoff’s schematic diagram of the modular structure of the “functional mind” is reproduced in Figure 4.1. Here, concepts are the fundamental semantic unit, the constituents of all thoughts, beliefs, desires, etc. The crucial point here is that there is no direct connection between conceptual structure and the world; instead, our interactions with the world are always mediated by our sensory apparatus, and linguistic
structures inherit meaning not from a direct relation to the world, but from a direct relation to conceptual structure.

[A] referent’s being in the world as conceptualized is a necessary condition for a speaker to refer. However, being in the real world is not a necessary condition: speakers can refer to entities like *Sherlock Holmes* and the *unicorn in my dream last night*. And an entity’s being in the world is not sufficient for reference either: one has to conceptualize it in at least some minimal way. In short, an entity’s being in the real world is neither a necessary nor a sufficient condition for a speaker’s being able to refer to it. Rather, the crucial factor is having conceptualized an entity of the proper sort. (Jackendoff, 2002, 304)

Jackendoff sees his analysis as in direct conflict with mainstream philosophical approaches to meaning. He quotes the likes of Lewis, Fodor, and Searle on this issue, each of whom emphasizes the need for truth conditions held accountable to an external world in order to get semantics off the ground. In fact, I think the disagreement here is largely one of rhetoric rather than substance. A closer look reveals that the basic idea behind Jackendoff’s strategy is not so different from those which have proved successful within the philosophical literature. These successful approaches also reject the straightforward identification of *that in virtue of which a term has meaning* with *that in virtue of which it refers to the external world*.

Let’s look at one particular criticism. In response to a similar position stated in Chomsky (2000), Fodor writes:

...in what way is the plausible claim that there are banks (that is, that there “really are” banks) not warranted? There is, for example, the bank that holds my mortgage. That is not just a way of talking; they make me pay up every month, cash on the barrel. How on earth could that be so if there really are not any banks at all? (Fodor, 2000, 4)

To which Jackendoff responds:

One reply would be that mortgages are just as much a mental construct as banks. But there is a deeper reply: we are ultimately concerned with *reality for us*, the world in which we lead our lives. Isn’t that enough?
... The reality in which you are reading this book and Jerry Fodor is paying his mortgage is certainly worth wanting; my claim here is that this world is a product of our human modes of perception and conception. If you want to go beyond that and demand a “more ultimate reality,” independent of human cognition, well, you are welcome to, but that doesn’t exactly render my enterprise here pointless. (Jackendoff, 2002, 309)

But Jackendoff goes way too far here. He doesn’t need to reject the external world or the existence of social institutions such as banks. All he needs is the claim that straightforward reference to such external structures as banks and mortgages is not adequate as an account of the meaning of terms like “bank” and “mortgage” (or the corresponding concepts). There’s no reason whatsoever why an internalist approach to semantics should provide a challenge to the existence of the external world. If the success of such a semantical approach is to count in favor of solipsism, an additional philosophical argument to this effect is required.

In fact, the idea that semantics and reference must be detached should not be surprising at all to philosophers, who themselves discovered the main problems which motivate this idea. Most notable, perhaps, is the problem of empty reference, which arises for expressions which contain terms referring to fictional or nonexistent entities, e.g. Jackendoff’s own examples above involving Sherlock Holmes and unicorns. Since at least the time of Frege, philosophers have acknowledged the challenge posed by these examples. Many recent strategies succeed at addressing these problems by detaching semantics from the “real” world in exactly the manner Jackendoff needs. I will briefly mention a couple examples in order to illustrate this point.

One strategy is exemplified by the neo-Meinongianism of Ed Zalta (e.g. 1993 and 1997). Zalta distinguishes between “existence” the traditional logical operator and “existence” in the sense of “having a location in spacetime” (Zalta, 1993, 403–4). The former notion of existence plays the usual quantificational role and identifies variables with objects in the model. The second notion of existence is treated as a distinguished predicate. By treating existence as a predicate, Zalta can save the traditional model-theoretic analysis of meaning along the lines of Tarski without having to worry about the fact that some of the objects to which terms in the language refer don’t exist in the actual world. Effectively, this strategy separates reference into two steps. The
semantics of a term is provided by its referent in the model. But the objects in the model are abstract—the question of whether they correspond to objects in the world (i.e. the existence predicate applies to them) is simply a separate issue.

A second strategy is exemplified by the holistic theory developed by Robert Brandom (1994 and 2000). Brandom wants to take discourse, and in particular the giving and taking of reasons, as more primitive than reference. He attempts to derive the meaning of a concept from its inferential role:

\[\ldots\text{for a response to have conceptual content is just for it to play a role in the inferential game of making claims and giving and asking for reasons. To grasp or understand such a concept is to have practical mastery over the inferences it is involved in—to know, in the practical sense of being able to distinguish (a kind of know-how), what follows from the applicability of a concept, and what it follows from. (Brandom, 2000, 48)}\]

Rather than beginning with the meanings of constituent terms and building up sentence meaning, Brandom begins with the inferential role of a sentence, then derives the meanings of terms from the intersection of all sentences in which they may occur. If the meanings of terms are derived from their inferential role, however, than they do not depend upon reference. In fact, Brandom’s own analysis of reference starts with the inferential role played by such technical terms as “reference” and “truth”, and a consequence of this analysis is that these terms cannot play the explanatory role they have been taken to play in traditional theories of meaning.

Understanding what it is to adopt such a representational stance toward \(\ldots\text{what we do is understanding the implicit practical attitudes that are expressed explicitly by our use of representational locutions} \ldots\) (Brandom, 1994, 280)

\[\text{[I]}\text{t is argued here that once the expressive function of ‘true’ and ‘refers’ is properly understood, it is seen to be incompatible with the explanatory function those locutions have been accorded in the dominant semantic theoretical tradition. (Brandom, 1994, 283)}\]

The strategies of Zalta and Brandom succeed at dealing with empty reference by postulating a semantics which does not depend upon reference. However, these
strategies do not satisfy the naturalistic desideratum that our semantics be derived entirely from a descriptive account of the world. Zalta must postulate a domain of abstract objects while Brandom appeals to inferential norms. Here, our strategy is to distinguish two distinct steps in the concrete flow of information from world to ideas. The picture here is broadly similar to the one in Perry (2001):

We learn about objects through perceptions. When we perceive things, we have ideas of them (notions), that we associate with ideas for the properties we perceive them to have, creating a sort of internal file (a notion associated with ideas). Thus the first link in our networks: perceptions of objects give rise to notions, and information flows from perception to notion. We retain these files when we are no longer perceiving the objects—the detach and recognize information game. We use language to share information that we gain in this way. . . . When we share information, the person doing the sharing constructs a statement with a reference in it, which is guided by the internal file he has of the objects. . . . The person receiving the information has a perception of the utterance. On the basis of it, he starts a new file or adds to one he already has. (Perry, 2001, 128–9)

Unlike the present work, Perry places himself within a broadly referentialist tradition. Nevertheless, his essential insight is consonant with the view developed here. The informational relationship between language and world is mediated by something like an “internal file”. My success at communicating with you depends upon a certain degree of agreement between our internal files. If I tell you “John’s car is red”, then your understanding of my claim depends upon the internal notions you associate with “John”, “car”, and “red”. Furthermore, we can distinguish this step, the association between words and internal folders, from the step which associates folders with the world. It is in virtue of the first step that words mean and communication succeeds; it is in virtue of the second that statements succeed or fail in making true claims.

Figure 4.2 illustrates the area in Jackendoff’s schematic on which I will focus. On this view, much as in Perry’s theory, the intermediary structure in virtue of which concepts acquire meaning is itself part of the mind. In the case of perceptual terms, this intermediary structure is the space of perceptual experience. The relationship A, from the external world to perceptual experience was analyzed as measurement
in the previous chapter. In the remainder of the present chapter, we will analyze the relationship between perceptual experience and detachable terms or concepts. The suggestion discussed in the following section is that concepts are derived from perceptual experience via tesselation, or the partitioning of a continuous space into discrete subspaces.

The upshot of this view is that the term “red” in the sentence “John’s car is red” will mean something in virtue of the relationship between the color term “red” and my perceptual experience of a range of values clustered in color space. So, the application of the predicate “is red” to John’s car will make a claim about the perceptual experience one would have upon perceiving John’s car, and will be true or false on the basis of whether that perceptual experience obtains.

Notice that this view need not collapse into dispositionalism about colors, the view that the property objects have when perceived as being some color is a disposition to produce that color experience under standard conditions. Whether or not an experience of redness obtains when we look at John’s car may depend upon whether or not the car is in fact red, i.e. participants in the color realism debate can accept the strategy taken here for detaching the meaning of color terms from the causal chain between world and experience without giving up on any particular analysis of what color is. Compare this point with Fodor’s example of the bank. Jackendoff can argue
that the term “bank” derives its semantic function from my internal concept of banks. This does not prevent Fodor from claiming that banks exist in the external world, and from analyzing them however he chooses (maybe a bank is a social construct, or a type of building, or a region in spacetime, whatever).

4.3 Gärdenfors’ Voronoi Tesselations

There are many different approaches one might take for deriving the semantics of perceptual concepts from perceptual experience. This section will outline a particularly suggestive strategy defended by Peter Gärdenfors (2000). I will first outline Gärdenfors’ general theory of conceptual structure, then highlight some of the details of his picture relevant to color and color terms.

Gärdenfors (2000) develops a theory of conceptual spaces. In previous chapters we have referred to spaces of perceptual experience, such as the color solid or the linear space of isolated pitch perception. These are essentially equivalent to Gärdenfors’ “conceptual spaces”, which are geometric structures with axes defined by continuously varying qualities. The color solid, for example, often has axes for hue, saturation, and lightness. For Gärdenfors, a “concept” is a region in a conceptual space, and it is from partitions of conceptual spaces that discrete mental symbols (such as terms in a language) gain their meaning. In Gärdenfors’ words, “to become meaningful, the symbolic level depends on the conceptual” (44).

Gärdenfors proposes the conceptual level to bridge the gap between the neural network level of “subsymbolic” representations and the more abstract level of “symbolic” representations. After a period of intensive debate between symbolic and connectionist camps in the late 1980’s and early ’90’s (initiated by Fodor and Pylyshyn (1988) and Smolensky (1987), though see Chalmers (1990) for an analysis of internal inconsistencies in Fodor and Pylyshyn’s argument), the dominant view that has emerged in cognitive science is a hybrid approach (see e.g. Anderson and Lebiere, 1998). Neural networks are good at learning while symbolic systems are good at modeling high-level cognition (language, belief, etc.), so why can’t they peacefully coexist together?

If neural networks and symbolic systems are to coexist, there must be principles
Figure 4.3: Network activation as vectors in a state space (Churchland and Sejnowski, 1992, Figure 3.3 (64), reprinted by permission of MIT Press)

for bridging them together. The problem here is that the activity of neural networks is naturally modeled by continuous spaces. The idea that an appropriate abstract characterization of the dynamics of a neural network is in terms of the movement of a vector through a high dimensional state space is introduced in the formal analysis section of Rumelhart and McClelland (1986). The basic idea here is that each node in the network defines a dimension (the degrees of activation it can exhibit), and the total state of a neural network can be modeled by a vector in an $n$-dimensional space, where $n$ is the number of nodes. As the neural network’s state of activation changes, the sequence of vectors characterizing it carves out a path in state space (Figure 4.3). A rousing empirical defense of this view is given in Churchland and Sejnowski (1992), while a more intuitive philosophical presentation can be found in Churchland (1995).

If neural network activity takes place in a continuous space, how can we associate it with the discrete space of symbols and concepts? Now, the present work does not take the reductionist view of the Churchlands, that perceptual experience reduces to a space of neuronal activation. Nevertheless, we have seen that perceptual experience is naturally modeled by continuous spaces, and we would like to understand the relationship between discrete spaces (such as the color terms) and these continuous perceptual spaces. Ideally, this relationship will provide a semantics for perceptual
CHAPTER 4. DISCRETIZATION AND DETACHABILITY

terms, and this is exactly the goal of Gärdenfors’ analysis.

Gärdenfors’ takes essentially the same attitude toward semantics as Jackendoff. His proposal is that conceptual spaces provide an abstract characterization of neural activity and a partition of such a space can provide the semantics for symbolic terms.

The core idea is that meanings of linguistic expressions are mental entities—meanings are elements of the cognitive structure in the heads of the language users . . . A semantics is described as a mapping from the expressions to a conceptual structure. This mapping can be seen as a set of associations between words and meanings—associations that have been established when the individual learned the language. According to this view, language represents a conceptual structure, but it does not directly represent the world. (Gärdenfors, 2000, 154)

Just as we have in the previous section, Gärdenfors distinguishes two steps in the relation between words and world. The first, semantic step, in virtue of which the words have meanings, is a connection between symbols and internal conceptual structures. The second step, from conceptual structure to the world is treated separately. Gärdenfors’ own approach is pragmatic:

The appropriate question on the relation between the conceptual structure and the world is whether a conceptual structure is viable or not. Having a viable conceptual structure means that one is able to solve the essential problems when acting in the world. Via successful interactions with the world, the conceptual structure of an individual will adapt to the structure of reality. (Gärdenfors, 2000, 156)

Unlike the present work, Gärdenfors insists that conceptual structure does not represent the world. Nevertheless, his basic picture is essentially consistent with the one developed here.

Gärdenfors’ strategy is to produce a partition of the continuous conceptual (or, in our terminology, perceptual) space into discrete subspaces; a procedure for inducing such a partition is called a tessellation. These discrete subspaces will then each constitute a “concept”, or the meaning of a particular term. Now, if subspaces could be produced arbitrarily, the theory would be vacuous, but Gärdenfors stipulates that
subspaces must be *convex* in order to be natural. In general, a space $Q$ is convex if for any two points $p, q \in Q$, all points in between $p$ and $q$ are also in $Q$; however, the appropriate notion of “betweenness” depends upon the type of space. If the space is Euclidean, then the points between $p$ and $q$ are just those which fall on the straight line segment which joins $p$ and $q$. Perceptual spaces need not, in general, be Euclidean, however (in fact, the best current theories of color space, our main example, is that it is non-Euclidean). So, it will be more accurate to state the definition of convexity in terms of geodesics, where a *geodesic* is just the shortest path between two points.

**convex set** – a set $Q$ is *convex* if for any two points $p, q \in Q$, all points which lie on the geodesic between $p$ and $q$ are also in $Q$.

Gärdenfors argues that only categories defined as convex subsets of a perceptual space should be treated as legitimate candidates for concepts. In support of this claim, he cites several empirical studies. We will be looking at relevant empirical research on color terms in the following section, but Gärdenfors also appeals to other perceptual domains. For example, Figure 4.4 reproduces data from Fairbanks and Grubb (1961). The two axes represent the pitch of the first two formants. A formant is just a peak in the amplitude of a voiced sound at a particular pitch caused by the shape of the vocal tract as the sound is uttered. They can be thought of as the most salient harmonics of a vowel sound, produced by the resonating properties of the human vocal tract.

Vowel discrimination depends upon the identification of the relationship between (at least) the first two formants. As can be seen in Figure 4.4 knowing the pitch of one of the formants is not enough. This is how humans identify vowels despite dramatic differences in the pitch of different voices. We can recognize an *ay* or an *ee* sound uttered at different pitches, because the relationship between base pitch (not depicted) and harmonics (first, second, third, etc. formants) are relatively stable. The complete details of this example are not relevant here. The essential point (which excites Gärdenfors) is that the regions associated with vowels in Figure 4.4 are all convex. Gärdenfors predicts that this property will generalize and that vowels in all languages will be identified with convex subspaces in formant space.
Figure 4.4: The frequency range of vowels graphed in the space defined by the first two formants. (Fairbanks and Grubb, 1961, Figure 1 (211), reproduced by permission of the American Speech-Language-Hearing Association)

As an aside, it is important to point out that the space depicted in Figure 4.4 is not actually an internal perceptual space such as the color solid. Rather, it is an external space defining the stimuli used in a psychophysical experiment. Although the regions identified as certain vowels have been graphed onto this space, it is a separate task to determine the internal perceptual space defining these experiences. In the language of the previous chapter, Figure 4.4 depicts the measured domain, not the measuring domain. However, this point of clarification is not meant to detract from the naturalness of Gärdenfors’ suggestion that identifiable concepts be restricted to convex subspaces within perceptual domains.

Convexity alone is not enough to partition a space, however; Gärdenfors also needs prototypes or clusters of prototypical points within the space. The clusters of points will define typical instances of a concept. Finally, Gärdenfors needs a metric on the space. The idea behind the voronoi tesselation is that one first solves for planes of discrimination between each pair of prototypes, then the union of these planes is taken to define the category boundaries around prototypes.
Suppose, for example, \( \mathbf{p} \) and \( \mathbf{q} \) are typical instances of two different concepts, i.e. each is a point within an \( n \)-dimensional perceptual space. This means that \( \mathbf{p} \) and \( \mathbf{q} \) are each vectors of \( n \) numbers defining coordinates within this space, i.e. \( \mathbf{p} = < p_1, p_2, \ldots, p_n > \) and \( \mathbf{q} = < q_1, q_2, \ldots, q_n > \). If the space is Euclidean, then the plane which discriminates between \( \mathbf{p} \) and \( \mathbf{q} \) is the solution to the equation

\[
\sum_i (p_i - x_i)^2 = \sum_i (q_i - x_i)^2
\]

Any \( \mathbf{x} = < x_1, x_2, \ldots, x_n > \) which is equidistant between \( \mathbf{p} \) and \( \mathbf{q} \) is a solution to this equation, so it defines the plane of all such points. The union of all such surfaces induces a partition on the perceptual space into discrete, convex subregions. This equation generalizes to non-Euclidean spaces in the obvious manner, by simply replacing Euclidean distance with the relevant measure of distance.

Notice that the arrow labeled \( \mathbf{B} \) in Figure 4.2 is bidirectional. There are well known top-down effects of categorization upon perception. Gärdenfors argues that the tesselation approach he endorses can explain such effects. For example, in the perception of phonemes, the physical stimulus may be varied continuously, yet the subject’s categorization behavior may demonstrate sharp boundaries. If one presents subjects with an artificially produced array of sounds varying continuously from a stereotypical \( \text{pa} \) to a stereotypical \( \text{ba} \), they tend to change their categorization of the stimuli quite abruptly despite the continuity in acoustical properties. The tesselation...
approach to analyzing this example has been implemented by Petitot (1989), who looked at a partitioning of the continuous space defining the articulation of stop consonants (b, p, t, d, g and k).\footnote{Petitot’s formal apparatus differs from that of Gärdenfors; in particular, he uses catastrophe theory to analyze nonlinearities in the acoustic signal, then uses these nonlinearities to derive the partition boundaries. However, the spirit of Petitot’s analysis and the data he collects clearly support the general spirit and claims of the Gärdenfors model.} On one axis is degree of voicing, on the other is locus of articulation (Figure 4.5). Petitot (and Gärdenfors) point out that the model generates empirically confirmed predictions, for example that “the contrast between /b/ and /t/ is much greater than that between /p/ and /d/” (because the categorical regions associated with /p/ and /d/ are adjacent while those associated with /b/ and /t/ are not) (Gärdenfors, 2000, 89–90; Petitot, 1989, 68).

Let’s pause to assess the plausibility of Gärdenfors’ model. On the plus side, it solves the problem of partitioning a continuous perceptual space into a discrete set of concepts with a simple algorithm. This is important because algorithms are mechanical, they may be implemented in our neural wiring, and, consequently, they satisfy our naturalistic desideratum. In fact, discriminating algorithms much like the voronoi tessellation are used in artificial intelligence as a means of categorizing complex data sets. In these applications, the programmer himself frequently does not know the appropriate category boundaries, so there is no danger of intentionality creeping in illegitimately.

On the down side, however, we are still quite far from actually implementing the research program Gärdenfors describes. It is telling that Gärdenfors emphasizes in his rhetoric the interest of providing an internalist semantics along the lines of Jackendoff; however, his empirical examples usually constitute partitions of the stimuli space, not the perceptual space. Both the vowel example in Figure 4.4 and the stopped consonant example in Figure 4.5 partition a space defined by properties of the stimuli. When I perceive a pa or a ba, these sounds fall within some continuous spectra of possibilities for me, but the locus of articulation in the mouth certainly does not define an axis of this space. Now, there may be an axis of variation in my perceptual experience which corresponds to the locus of articulation, this is an empirical question. In order to think clearly about perceptual representation, however, we need to keep this internal
dimension distinct from the external one it measures. It should be clear that the locus of articulation is what is being measured, not what is measuring in my experience.

The reason for this discrepancy is perhaps just the difficulty in rigorously defining perceptual spaces. As we have seen in our discussion of color, not the least of these difficulties is in determining the appropriate metric. In the case of physically defined stimuli, distances are easy to come by, but in the psychophysical investigation of experience, they can be very difficult to determine. Now, for something like voiced vs. unvoiced, or even for a range of first formant pitches, it is perhaps relatively safe to assume that some axis of some level of perceptual experience corresponds to these axes. Not so for a space like color, however, where the dimensionality of the perceptual space is significantly smaller than that of the external spaces being measured.

In the next section we will examine empirical evidence in favor of an internalist semantics of color terms along the lines suggested by Gärdenfors. I would like to close this section with a very suggestive application of the Gärdenfors strategy, however.
In Chapter 2, we briefly discussed the application of basic color terms to human skin, whether in racial classifications or supposedly “metaphorical” locutions like “green with envy”. Gärdenfors proposes that the semantics for such applications can be provided by restricting the space of possible color experience to just that subspace within which possible human skin tones fall. Then, color terms apply to regions within this subspace which correspond to the regions of their application in the overall color solid.

I want to show here how the idea that a contrast class determines a domain can quite easily be given a general interpretation with the aid of conceptual spaces. For each contrast class—for example, skin color—one can map out the possible colors on the color spindle. This mapping will determine a subset of the full color space. The shape of this subset may be rather irregular. Now, if the subset is completed to a space with the same geometry\(^2\) as the full color space, one obtains a picture that looks like [Figure 4.6].

In this smaller spindle, the color words are then used in the same way as in the full space, even if the hues of the color in the smaller space do not match the hues of the complete space. Thus, “white” is used about the lightest forms of skin, even though white skin is beige, “black” refers to the darkest form of skin, even though black skin is brown, and so forth. Note that the set of possible skin colors will not cover all of the small spindle, but certain skin color regions will be empty. There are, for example, no green people (but one can become green with envy or because of sickness). (Gärdenfors, 2000, 121–122)

Now, this is a suggestion which requires empirical support. To my knowledge, there is no study which has attempted to isolate the exact subspace of the color solid which applies to human skin tones, so some work needs to be done here to verify this analysis. Nevertheless, the rough intuitive strategy here is quite suggestive, and applies to many domains in which contrastive terms are used (e.g. hot / cold, tall / short, loud / soft, funny / serious, etc.). The idea that a building is tall if it falls at one end of the spectrum of possible building heights, while a person is tall if he falls at one end of the spectrum of possible people heights, is completely plausible.

\(^2\)This restriction would have to be weakened in a realistic application of this idea. Since color space is not isotropic, it will not in general be possible to define a subspace with exactly the “same geometry” as the full space.
4.4 Recent Empirical Evidence for the Gärdenfors Model

Here is a hypothesis about color terms: color terms derive their meaning from the color solid (the space of perceptual color experience) by partitioning it into convex subspaces. How can we turn this into an empirical claim? Let’s remind ourselves of the strategy discussed in the previous chapter for determining what a representational system represents. Examine a number of different systems which presumably represent the same thing and analyze the invariants across permissible mappings between them. In the terminology of Chapter 3: look for invariants across $\beta$-mappings and use these as evidence for the nature of $\alpha$-mappings (see Figure 4.8).

In the case of color terms, we can investigate the color terms found in different languages and compare their domains of application. Hopefully, some convergence amongst these domains will provide evidence for the correct semantic analysis of color terms. The seminal work in such a project is Berlin and Kay (1999), and the basic research program they outline and defend is continued in Hardin and Maffi (1997) and the World Color Survey.$^3$

Berlin and Kay and the World Color Survey collected data about color terms in two stages. The first stage is identification of the basic color terms in a language. This stage is done through analysis of the distribution and application of terms in a target language. In order to count as a basic color term, for example, a candidate must be monolexemic (i.e. the meaning cannot be derived from parts of the term, ruling out e.g. “blue-green”); its application must not be restricted to a narrow class of objects (ruling out e.g. “blond”); and it may not be a refinement of a more general term (ruling out e.g. “crimson”). These and a number of other specific criteria are discussed in Berlin and Kay (1999), Section 1.2.

In the second stage, native speakers are asked to identify the most representative instances of a color term and the boundaries of its corresponding category on a Munsell color chart. This chart contains 330 systematically arranged color patches. 10 of

$^3$A monograph assessing the data of the World Color Survey is currently in preparation. Their data are available online at http://www.icsi.berkeley.edu/wcs/data.html.
these patches depict a range of achromatic greys from black to white. The remaining 320 patches are arranged in an $8 \times 40$ grid, with each column corresponding to a unique hue and each row corresponding to one of eight degrees of brightness. These patches are produced in accordance with the Munsell Book of Color, one of the most popular and widespread systems for organizing color percepts.

The Munsell color system is one attempt to organize color percepts in terms of a relevant distance metric. Recall the discussion at the end of Chapter 2 where we emphasized that distances in a tristimulus coordinate system derived from color matching experiments do not correspond to subjective assessments of the distances between color experiences. The Munsell system partially corrects for this deficiency by attempting to organize color samples in a spacing which corresponds to equal perceptual distances. In the late 1930’s and early ’40’s the Munsell Color Company worked closely with the Optical Society of America to achieve this end. The corresponding system is not isotropic and distances across hues or brightness levels are not meaningful. However, within a given hue, lightness distances approximately correspond to degrees of perceptual distance, and at a fixed brightness, hue distances are roughly equal to each other from chip to chip (see e.g. Kuehni and Schwarz (2008), 160–1).

So, the stimulus used for the World Color Survey can be thought of as a flattened cylinder, with the circular axis denoting hue and the vertical axis denoting lightness (Figure 4.7). Berlin and Kay (1999) collects data on how color terms apply to these color patches for 98 languages; the World Color Survey has been expanding this data
set to include many additional languages.

In 1969, Berlin and Kay drew two conclusions from their data. The first took a stance in the debate about linguistic relativism. The relativistic hypothesis is that the assignment of color terms is purely conventional, and thus carves up color experience in an essentially arbitrary fashion. If this were the case, then there should be no substantial agreement across diverse languages about either the location of focal colors or the boundaries of color categories. But Berlin and Kay’s crosslinguistic survey found dramatic agreement in the foci of color terms:

> Our results . . . cast doubt on the commonly held belief that each language segments the three-dimensional color continuum arbitrarily and independently of each other language. It appears now that, although different languages encode in their vocabularies different numbers of basic color categories, a total universal inventory of exactly eleven basic color categories exists from which the eleven or fewer basic color terms of any given language are always drawn. The eleven basic color categories are *white*, *black*, *red*, *green*, *yellow*, *blue*, *brown*, *purple*, *pink*, *orange*, and *grey*. (Berlin and Kay, 1999, 2)

Although Berlin and Kay set out to contribute to the linguistic relativism debate, they came across their second finding purely by accident. Once they began comparing terms across languages, it became apparent that there was a fixed ordering to these systems. Not only do languages range from those with only two basic color terms to those with eleven, but there appear to be basic constraints on the locations of foci for color terms which depend upon the number of basic terms. On the basis of this observation Berlin and Kay hypothesized an evolutionary progression from fewer to more color terms, with constraints on the order in which terms can be acquired.

Figure 4.8 depicts the relationship between some languages with different stages of color terms in the Berlin and Kay hierarchy. English is an example of a “Stage VII” system with 11 basic color terms (white, black, red, green, yellow, blue, brown, purple, pink, orange, and grey). Plains Tamil is a typical “Stage V” system found in South India, with 6 basic color terms (white, black, red, green, yellow, and blue). Tiv is Bantoid language found in Nigeria, it is a typical “Stage II” system with 3 color terms (white, black, and red). Jalé is found in the highlands of New Guinea and is classified
as a “Stage I” system with only 2 color terms (white and black). The use of the same basic color terms to characterize the systems of each of these languages is legitimated by the close clustering of focal points on the Munsell color chart for terms across languages. Since this clustering holds even across languages with different numbers of color terms, the $\beta$-mappings here represent homomorphisms which collapse nearby color categories into larger groupings.

Assessment so far: If we accept the results of Berlin and Kay (1999), we see partial but ambiguous evidence for the internal semantics story. Berlin and Kay argue forcefully for universalism over relativism about colors. So, if we accept their arguments, we accept that color categories are not merely social constructs or arbitrary classifications of surfaces. However, the source of the agreement here is unclear. Do the foci of color terms in different languages converge because of some property of our perceptual system or because of the fixity of structures in the external world which colors represent? In Figure 4.8, that aspect of color term structure which is preserved across $\beta$-mappings is the focal points of color terms. So, by the arguments in Chapter 3, the focal points for color terms are meaningful. This isn’t enough, however, to determine what is meaningful about them. It could be that the physiology of color vision determines the meaningfulness of focal points in color space, in which case the internalist semantics discussed in the previous sections would be vindicated. However, it could just as easily be external properties associated with these focal points (redness, greenness, etc. out in the world) which ensure their invariance across diverse languages. Why do our color terms agree about the shade of focal red? Perhaps just

Figure 4.8: $\beta$-mappings amongst systems of color terms.
CHAPTER 4. DISCRETIZATION AND DETACHABILITY

because they represent the reflectance properties associated with that shade.4

Despite their compelling data, the methodology of Berlin and Kay has been challenged repeatedly. The most famous challenge is perhaps Lucy (1997) (though see also Saunders and van Bradel (1997) and the response in Kay and Berlin (1997)). Lucy emphasizes the results of Conklin (1955) on Hanunóo as well as the work of Lenneberg and Roberts (1956) and Hickerson (1975) on Zuni. In Chapter 2 we discussed how “color” terms in Hanunóo appear to refer not only to hue, but also to other features of the surface, such as freshness vs. desiccation. In Zuni, “color” terms distinguish not only hue but how the surface hue came to be. For example, different terms for “yellow” mark whether the surface was made to have that hue (as through painting) or whether that hue arose through a natural process (e.g. the yellowing of leaves).

Lucy argues rather forcefully that the work of the World Color Survey is not good linguistics. This is because no work has been done from a linguistic standpoint to verify that “color terms” constitute natural categories within any of the languages under investigation. The traditional linguistic method for determining natural categories of terms is to look at their distribution. But even in English, the so-called “basic color terms” do not constitute a natural category by this criterion. For example, although yellow and pink are both supposed basic color terms, they admit of different forms—yellowing is possible while pinking* is not. Neither can take an –en suffix as in reddening (Lucy, 1997, 329). Lucy concludes that the World Color Survey has not demonstrated adequately that it investigates a legitimate linguistic category, and thus cannot claim to contribute to our understanding of the semantics of that category.

Do Lucy’s criticisms defeat the evidential value of the World Color Survey data for determining the semantics of color terms? Although they indicate crucial weaknesses in the basic methodology, esp. at the first stage, in which basic color terms are identified, they do not completely defeat the project. Berlin and Kay are well aware of the features discovered by Conklin (1955) and Lenneberg and Roberts (1956), yet

4I cannot find Berlin and Kay’s study used in this fashion to support color realism anywhere in that literature. Nevertheless, this seems an obvious argument to make given the discussion of representation in the present work.
emphasize that as long as hue and lightness also contribute to the reference of the relevant terms, the methodology of the Color Survey is legitimate:

[I]t has been argued, to our minds convincingly, that to appreciate the full cultural significance of color words it is necessary to appreciate the full range of meanings, both referential and connotative, and not restrict oneself arbitrarily to hue, saturation, and brightness. We thus make no claim—in fact we specifically deny—that our treatment of the various color terminologies presented here is an ethnographically revealing one.

The data presented in this monograph are admittedly removed from their cultural context; however, we can not accept the stricture offered by some ethnographers that such removal always and necessarily renders data meaningless. The high degree of pattern found in the data is sufficient justification for the process dictating its selection. We thus interpret the pattern found in our results as representing legitimate linguistic and cultural universals. (Berlin and Kay, 1999, 160)

Indeed, the convergence in the data from the World Color Survey seems to demand explanation despite the worries expressed by Lucy. However, other critics have worried that the convergence in the data is due to cultural bias inherent in the materials used (i.e. the Munsell chart and the method of presenting it to the subjects). This is the line taken by Saunders and van Bradel (1997), and rather forcefully rebuffed by Kay and Berlin (1997). Critics such as Saunders and van Bradel have failed to convincingly identify any particular feature of the methodology of the World Color Survey which introduces such bias. (Short of a thoroughgoing postmodernism which demands relativism about all scientific claims.)

However, there is a fly in the ointment. The universalist hypothesis seems supported by the convergence of focal points for color terms across languages. Unfortunately, the boundaries of color categories do not converge in the same manner. The degree of variation in the boundaries of color categories across languages has given relativists ammunition against the Berlin and Kay hypotheses. The most recent step in this debate tempers the universalist hypothesis and provides a new explanation for the patterns found in the data of the World Color Survey. This new explanation also provides explicit support for an internalist semantics of color.
The basic suggestion is that the convergence in color categories across languages is derived from asymmetries in the color solid. A “good” color naming system should partition the color solid efficiently and the constraints on such efficient partitions provided by the asymmetrical structure of the color solid is the source of the convergence of color naming systems across languages. If this analysis is correct, it supports internalist semantics because the content of color terms is identified with partitions of the space of color experience rather than some more distal feature (such as surface properties in the world).

This suggestion was originally presented as a criticism of the World Color Survey by Jameson and D’Andrade (1997); it has recently been formalized and empirically confirmed by Regier et al. (2007). The first step in formalizing the claim is to decide on a model for the space of perceptual color experience. As discussed above, the Munsell system is used to organize the stimuli for the World Color Survey, but it does not define meaningful distances between samples which differ with respect to both hue and brightness. Regier, et al. begin by mapping the color samples used in the Survey into the CIELAB color space (Figure 4.9).
CIELAB is a color space developed by the Commission internationale de l’éclairage in 1976. It has become the universal standard for colorimetry despite known weaknesses. Like many color spaces, CIELAB models color experience with a roughly spindle-shaped space. The vertical dimension $L^*$ corresponds to lightness and darkness, while the $a^*$ dimension roughly corresponds to red / green and the $b^*$ dimension roughly corresponds to blue / yellow. As a color appearance model, there are many shortcomings to CIELAB, such as known features of color experience which it cannot adequately capture (see discussion in Fairchild, 2005, Chapter 10).

For the purposes of understanding the WCS data, however, CIELAB is a natural choice. Its limitations derive from the fact that it is intended only to model the perceptual experience of stimuli of uniform size presented against a neutral background. This means that there are contextual effects it cannot predict. However, the Munsell stimuli are of uniform size and presented against a neutral background. So, none of the data collected for the WCS should exhibit the effects which CIELAB fails to model. Furthermore, CIELAB offers the crucial advantage over the Munsell system that distances are meaningful across changes in hue, saturation, and lightness.

The nonuniformities in the Munsell stimuli are immediately apparent when they are mapped into CIELAB color space. Remember that the stimuli used by the WCS are intended to be of uniform saturation. If color space were completely homogenous, we would expect these samples to fall on the surface of a cylinder. Instead, however, we see in Figure 4.9 that they produce a distorted spheroid. As discussed previously, there is a significant bulge in the yellow region (in the positive direction along the $b^*$ axis)—the wider spacing of samples here indicates our ability to perceive more fine-grained distinctions in the red–green region of the spectrum than in the blue and purple regions. This deformity derives from the overlap in the spectral receptivity of $L$ and $M$ cones in the retina. The second most noticeable bulge occurs in the low brightness region of the purple area of the spectrum. Again, this is to be expected given our knowledge of the physiology of color vision. Wavelengths perceived as purple fall at the outer reaches of the visible region of the spectrum; correspondingly, they produce weaker signals on average in the cone cells, resulting in the purple hues falling largely in the low brightness (here, low $L^*$) region of color space.
The next step in the analysis of Regier et al. (2007) is to produce a plausible criterion by which to evaluate the efficiency of a color naming system.

We wish to characterize how good a categorical partition of color space this arrangement represents. To that end, we defined an objective function that measures the extent to which such an assignment of category labels to chips maximizes similarity within categories and minimizes similarity across categories. We refer to this quantity as “well-formedness”: optimal partitions of color space are those that maximize this well-formedness measure. (1438)

Since CIELAB comes with a meaningful distance metric, similarity can be cashed out in terms of some function on this metric. In particular, Regier et al. use

$$\text{sim}(x, y) = \exp(-c \times \text{dist}(x, y)^2)$$

where dist is just the CIELAB distance metric and c is a scaling factor. This measure of similarity assigns 1 if \(x = y\) and some fraction which drops off toward zero as the distance between \(x\) and \(y\) increases. This similarity measure is then used to define a measure of well-formedness \(W\).

$$S_w = \sum_{(x,y): \text{cat}(x) = \text{cat}(y)} \text{sim}(x, y)$$

$$D_a = \sum_{(x,y): \text{cat}(x) \neq \text{cat}(y)} (1 - \text{sim}(x, y))$$

$$W = S_w + D_a$$

Where \(\text{cat}(x)\) is just the color category in which \(x\) is placed. Here, \(S_w\) measures the degree of similarity amongst category members by summing over their pairwise similarity distances. \(D_a\) measures the degree of dissimilarity across category boundaries (more dissimilarity between \(x\) and \(y\) implies a lower value for \(\text{sim}(x, y)\) and thus a higher value for \(1 - \text{sim}(x, y)\)). \(W\) simply sums these two measures together (Regier et al., 2007, 1438).
Regier et al. demonstrate two results with their well-formedness measure. The first suggests that optimal (i.e. maximally well-formed) partitions of color space are found in natural language color categories. The second suggests that all systems of color terms are in some sense optimal.

The first result requires the artificial generation of a maximally well-formed categorization of color samples. This can then be compared with attested color naming systems to see if any partition color space in the same way. Regier et al. produced optimal color categorizations for systems with 3, 4, 5, and 6 distinct color categories. The procedure begins with a random assignment of samples to color categories. The categorization of samples is then changed one by one ensuring that each change increases the value of $W$. Eventually a local maxima is reached and an optimal distribution is found. For each number of categories, this procedure was implemented 20 times, then the categorization amongst these 20 which received the highest value for $W$ was compared to attested categorization schemes with the same number of categories.

For each artificially generated optimal categorization, several naturally occurring color naming systems could be found which matched it very closely. This is not to say the match was ever exact, and in some cases, there were systematic discrepancies. For example,

This deviation is especially pronounced in the blue region. Starting with three-term languages and continuing through six-term languages, the category that includes green in the theoretically optimal configurations usually does not extend far enough “rightward” into blue/purple . . . Thus, with regard to blue, this model makes the wrong prediction. (1438)

However, these systematic discrepancies are relatively minor, and may reflect inaccuracies within the CIELAB model rather than a failure of the efficiency hypothesis.

The second result requires some test of the optimality of attested color naming schemes. As Regier, et al. acknowledge, the details of a particular set of color terms will depend upon the history of that language. As such, it is not unsurprising that many naturally occurring systems of color terms do not match the artificially generated optimal categorizations just discussed. Nevertheless, by hypothesis they should
Figure 4.10: Well-formedness of Berinmo color categories when rotated around the color solid (Regier et al., 2007, Figure 6 (1440), reprinted by permission, copyright (2007) National Academy of Sciences, U.S.A.).

exhibit some kind of local maximum efficiency in their partitioning of color space. In order to test this hypothesis, Regier et al. took attested color naming systems and rotated their categories around the cylinder of Munsell stimuli (Figure 4.7). Since the $8 \times 40$ grid of stimuli is completely symmetrical, this rotation is well defined: relative category boundaries were preserved, just shifted over by two columns at a time. Since the mapping of these chips into CIELAB space is asymmetrical, however, the value of $W$ should change with each rotation. If a system of color terms under investigation constitutes a local maximum in the space of efficient partitions of color space, then any rotation of the system away from its original orientation should decrease its well-formedness $W$.

Regier et al. performed this procedure for all 110 languages in the WCS data. Of these, 82 languages exhibited maximal well-formedness with no rotation. Of those for which well-formedness increased with rotation, most maximal values for well-formedness occur with very little rotation away from the zero point. Regier et al. present the data for Berinmo in detail. Figure 4.10 clearly demonstrates how the well-formedness of the Berinmo categorization system drops off dramatically as it is rotated further around the color solid.

Berinmo was chosen for close discussion because it has been suggested recently
as a counterexample to the supposed convergence of focal points for color categories (Roberson et al., 2000). The discussion by Roberson, et al. does not undermine the position defended here, however. They demonstrate the effects of color terms on memory and categorization tasks. But these results support the general perspective of internalist semantics. The color solid (a model of our perceptual experience) derives from physiology and is thus as universal as the physiology of color vision. Color terms, in contrast, are detachable from this color solid. Regier et al. demonstrate that the only universality in color naming schemes is in the efficiency with which they partition the color solid. But this universality supports the general claim that color terms derive their meanings from the color solid, not from more distal features in the causal chain between experience and world such as surface properties.

4.5 Conclusion

We began this chapter by discussing the internalist approach to semantics. On this perspective, the problem of assigning meanings to terms is distinct from the problem of determining how terms connect to the external world. Meanings depend upon the relation between lexemes and our internal model of the world. Our internal models of the world must connect somehow with the external world in order to help us navigate reality. However, this relationship, between internal model and external world, is distinct from the internal relation in virtue of which terms have meanings. If this internal relationship can be naturalistically identified, then the problem of detachability has been solved. This is because such a relationship constitutes a naturalistic semantics for linguistic items which does not depend upon the presence of their referents for the terms to be tokened.

There is no complete worked out version of internalist semantics along these lines. However, Gärdenfors has implemented an approach which appears to work well for terms and concepts closely related to perceptual experience. His approach uses a tessellation algorithm to partition continuous spaces of perceptual experience into discrete subspaces, each of which provides the semantics for a conceptual category. This strategy satisfies our naturalistic desiderata since it invokes an algorithm which
may be implemented on any machine. Thus, Gärdenfors’ theory makes an empirical claim: either our brain does or does not compute partitions of perceptual spaces.

Next we examined a recent result in detail which appears to support the internalist approach to the semantics of color terms. The World Color Survey investigates how the color terms of different languages partition perceptual color space. Some features of the data seemed to indicate a convergence of color categories across languages. This result is ambiguous for an internalist semantics of color: this convergence may be due to universals of perceptual experience or it may be due to regularities governing the referents of color terms in the external world. However, more recent analysis indicates that the only real universal across the color naming systems of different languages is the efficiency with which they partition perceptual color space. This strongly suggests that it is not distal referents, but the more proximal relationship between color terms and color experience from which color terms derive meaning.

Finally, this analysis can be seen as an instance of the general research strategy outlined in the previous chapter. We have examined invariances across different color naming schemes and used these invariances as evidence for what is being represented by the color terms. Since these invariants depend upon the structure of perceptual color space, not the physical structure of color stimuli, they speak strongly in favor of the idea that color terms are about color experience, and are used to refer to surfaces, lights, etc. in our environment only in a derivative sense.
Chapter 5

The Informational Content of Natural Signs

5.1 Introduction

So far, we have focused on how error can emerge in natural representational systems. We saw how any statistical interaction (such as that which occurs between photons and photopigments at the retina) induces imprecision. Next we saw that relative inaccuracies can be discovered between different representational subdomains which measure the same external domain. Here, we have no problem determining that both spaces measure the same feature of the world because the stimuli used and the experiences reported upon are identical in each case (as in Weber’s experiments on weight). The difference comes merely in what informationally relevant features of sensation we control for in our experimental design (e.g. temperature in Weber’s Thaler experiment).

Finally, in the last chapter we saw how representational structures could be detached from their referents in the external world if we allow for a two-stage semantics. The first stage is the informational relationship between our experience and the world (analyzed as a measurement of the world). The second stage is the semantic relationship between discrete terms or concepts and the continuous space of perceptual experience. This two stage structure allows for referring structures such as words to
be detached from the causal chain between experience and world.

However plausible this story may seem, two key points demand a thorough discussion. First, Grice (1957) has argued that natural meaning can never be in error. So far, we have attempted to demonstrate exactly how meaning in the natural world can be susceptible to error. Perhaps, however, we have gone astray somewhere; if the impossibility of error is constitutive of natural meaning then the whole artifice constructed here rests on sand. I will argue in Section 5.2 that Grice’s argument has been misunderstood by the literature on naturalizing meaning. Once we realize that the relationship of natural signification holds between types rather than tokens, we see that Grice’s considerations simply do not apply to the present analysis.

Second, this leaves us with the question of what exactly natural meaning is. I have repeatedly used the expression “informational content”, but what exactly do I mean by this? Measurement outcomes by themselves are not enough because any naturally varying space measures many different domains of variance. We need some way to combine all these measurement outcomes into a single formal object, and information theory suggests such a strategy. Section 5.4 will introduce a probabilistic analysis of informational content derived from Shannon information theory. This analysis is meant as an alternative to Dretske’s notion of “semantic information”, which rejects the probabilistic apparatus of information theory. I believe this rejection was motivated by two factors. First, Grice’s argument, which we will have dispensed with by that point. Second, the frequently expressed worry that information theory itself does not provide an analysis of informational content. Here, however, Dretske’s understanding is just outdated and a close examination of recent applications of information theory reveals a helpful analysis of informational content.

I will call the crucial structure here an $s$-vector. $s$-vectors organize the complete informational content of an event into a single formal object. After introducing enough of the apparatus of information theory to explain $s$-vectors, we will look at both theoretical motivations (most recently due to Skyrms) and empirical applications. It is telling that the same analysis of content has emerged independently from a priori and empirical considerations.
5.2 The Problem of Error, part 3: Natural Signs and Natural Meaning

In a seminal paper, H. P. Grice (1957) argues that there are two distinct ways we use “meaning”, which he dubs “natural” and “non-natural”. Consider, for example, these two uses of the word “mean”:

1. “Those spots mean measles.”

2. “Those three rings on the bell mean the bus is full.”

A crucial difference between the use of “mean” in sentences 1 and 2 is its factivity. In Grice’s words “I cannot say, ‘Those spots meant measles, but he hadn’t got measles’” (377) yet I can utter 2 “and go on to say, ‘But it isn’t in fact full—the conductor has made a mistake’” (378). Grice dubs the first, factive use of “mean” natural meaning, and the second, non-factive use non-natural meaning. His analysis further distinguishes these two senses of “meaning” by noting additional differences in their distribution. From a semantic standpoint, however, these additional distributional differences appear tied to the essential insight, namely that whenever “meaning” is used in the natural sense

\[ \ldots x \text{ meant that } p \text{ and } x \text{ means that } p \text{ entail } p. \] (Grice, 1957, 377)

Now, there is a long philosophical tradition investigating a kind of natural meaning before Grice. This tradition focuses on the notion of a “natural sign”, an idea which dates at least to 396 AD and Augustine’s De Doctrina Christiana:

Natural [signs] are those which have the effect of making something else known, without there being any desire or intention of signifying, as for example smoke signifying fire. It does not do this, after all, because it wishes to signify; but through our experience of things and our observation and memory, we know that fire is there, even if only smoke can be seen. (Augustine, 1996, 129)
Augustine’s *signa naturalia* are signs ready to be interpreted; each is a word in the language of nature waiting to be read. Classic examples of natural signs (smoke as a sign of fire, clouds as a sign of rain, footprints as a sign of prey, etc.) clearly fall within the category of Grice’s natural meaning if uttered with “mean” in between. I can say “these clouds mean rain” and “these footprints mean dinner”, but if I immediately follow that with an assertion that it won’t rain, or there won’t be any dinner, I seem to have said something funny or counterintuitive.

Grice (1957) never uses the term “natural sign”, but Dretske and Millikan both take him to have demonstrated an essential feature of natural signs. They cite Grice’s argument as demonstrating that natural signs must be factive—clouds do not mean rain *unless it rains*; tracks do not mean quail *unless they were made by a quail*:

As Grice observes, nothing can mean that *P* in the *natural* sense of meaning if *P* is not the case. This distinguishes it from non-natural meaning. ... A person can *say*, and *mean*, that a quail was here without a quail’s having been here. But the tracks in the snow cannot mean (in this natural sense of “meaning”) that a quail was here unless, in fact, a quail *was* here. (Dretske, 1988, 55–6)

So, Dretske endorses Grice’s point about the factivity of meaning, but he transposes it from discourse to nature. Not only are we not allowed to *say* “these tracks mean quail, but they were not made by a quail”, but tracks themselves do not signify quail if they were not made by quail.

But there is something funny going on here. Natural signs are the sort of thing we can come along and use to learn about the world. If I see smoke, I can learn about the presence of fire; if I see clouds, I can learn about the presence of rain. A theory which stipulates that clouds don’t mean rain unless it in fact rains can’t support my use of clouds to learn about rain.

This is because the category I identify in nature is not the set of clouds which are followed by rain; it cannot be defined in terms of the future presence of the signified. Rather, I identify a category of items which appear similar to me, call them *storm-clouds*. Storm-clouds all share salient features which I can identify (maybe they are low, thick, and dark). Yet weather is chaotic; the lowness, thickness, and darkness of
a cloud alone is not enough to ensure that it is followed by rain. This poses a problem for the application of Grice’s factivity constraint to natural signs.

Dretske and Millikan both recognize this issue, but deal with it in different ways. Dretske’s response is just to strengthen his definition of natural sign, stipulating that in any region within which the potential signifier may occur without the potential signified, there is no natural signification.

...Furthermore, even if \( P \) [the signified] does obtain, the indicator or sign does not mean (indicate) that \( P \) is the case unless the requisite dependency exists between the sign and \( P \). Even if the tracks in the snow were left by a quail, the tracks may not mean or indicate that this is so. If pheasants, also in the woods, leave the very same kind of tracks, then the tracks, though made by a quail, do not indicate that it was a quail that made them. (Dretske, 1988, 56)

If we apply this reasoning to the storm-cloud example, we get something like this: storm-clouds mean rain within some (spatiotemporal) region if and only if there are no clouds which are identical to storm-clouds but not followed by rain within that region. So, Dretske acknowledges the possibility that a member of a signifying natural category may not stand in the appropriate causal relationship to members of a signified natural category. “Quail-like” tracks may not be made by quail; “storm”-clouds may not be followed by rain. However, he simply denies that a significational relationship obtains if any quail-track-producing pheasant or non-rain-producing storm-cloud appears within the relevant region.

Millikan takes a different approach to the same problem. She worries that Dretske’s approach does not take sufficiently seriously the actual environments within which natural signs come to be used. In particular, she correctly points out that Dretske’s stipulation that informational success depends upon the signified obtaining with probability one, conditional on the signifier, will never hold for his own examples:

But surely, whether or not there are pheasants as well as quail in these particular woods is a matter of statistical frequency, not natural law. If we look only at natural law, given that pheasants as well as quail could lawfully produce such tracks, there could never be a probability of one that such tracks are produced by quail, no matter how far these tracks
happened to be from any actual pheasants. Logic and natural law do not change over space and time with variations in the distribution of pheasants and quail. (Millikan, 2004, 32)

Millikan concludes

...that a theory of natural information that will help to explain how real animals manage to obtain useful information will need to introduce statistical considerations about the environment in some controlled way. (Millikan, 2004, 34)

However, despite this assertion, Millikan’s ultimate analysis of natural signs retains the condition of factivity. She at first appears to give it up by emphasizing that Dretske must replace his insistence on law-like relationships as the basis for natural meaning with correlation. Natural signs (in Millikan’s terms, “locally recurrent signs”) are then those for which correlation holds within a “natural domain”:

A natural reference class for a sign—the natural domain within which certain As are “locally recurrent signs” of certain Bs—is a domain within which the correlation of As with Bs extends from one part of the domain to other parts for a reason, and it must be a domain that it is possible for an organism to track. (Millikan, 2004, 40)

But what are “natural domains”? A natural reading of Millikan’s discussion is that they are spatiotemporal regions, e.g. our forest. These tracks are a sign of quail within the forest only if the correlation of quail presence with tracks extends from one part to the other. This type of interpretation has been developed by, for example, Shea (2007).

However, in subsequent works, Millikan has explicitly rejected this interpretation:

The confusing point is that the “domains” of locally recurrent signs are not the same as areas in which they correlate with their representeds, unclarity or equivocation on this point having muddied some fairly central passages in Varieties of Meaning. Instead, the “domain” of a locally recurrent natural sign is like the domain of a function or quantifier. It is merely the set of all actual instances falling under that locally recurring sign type. Since “areas” can be of any shape whatever, including shapes
with numerous irregularly shaped holes in them, the domain of a recurrent natural sign type does not determine any definite area. On the other hand, the domain does help determine, for any given area, the proportion of signs of the same physical type that are also within that domain. Although its area of residence does not determine whether a physical sign does or does not fall in a given locally recurrent sign domain, locally recurrent sign domains do help determine statistics on already given areas. Sometimes organisms just happen to live and die within areas where the statistics on a certain recurrent sign are good or good enough. Other times they may develop crude or less crude ways of tracking locally recurrent sign domains well enough to be useful—ways of tracking that work, at least, in the areas in which they live. (Millikan, 2007, 453–4)

So, for Millikan, there are two kinds of storm-clouds, call them clouds$_1$ and clouds$_2$. Members of clouds$_1$ and clouds$_2$ are physically indistinguishable, i.e. they are “of the same physical type”. However, after members of clouds$_1$ there is rain and after members of clouds$_2$, there is not. For Millikan, natural signification is like a function from members of clouds$_1$ to instances of rain; in this sense, clouds$_1$ is the “domain” of the natural sign linking clouds to rain. So, the relationship of signification is factive: only members of clouds$_1$, only clouds following which it then rains, are signs of rain. But the ease with which members of clouds$_1$ can be used as signs of rain depends upon the ability of potential users to “track” that “domain”, which in turn depends upon the area within which the users dwell. In our forest, for example, the correlation between the physical type clouds$_1$ $\cup$ clouds$_2$ and rain may be very high because there happen to be many more instances of clouds$_1$ obtaining in the forest than of clouds$_2$.

Neither attempt to analyze natural signs is satisfactory. Dretske’s analysis fails because the situation he describes never (or hardly ever) obtains. Certainly, his account cannot explain typical examples of natural signs because the causal relationship between signifier and signified is too convoluted to ensure Dretske’s condition that signifier $iff$ signified ever be satisfied. (See Appendix A for a discussion of the causal relationship between smoke and fire and between spots and measles.)

Millikan’s account fails for two reasons. First, one might argue it violates the minimal naturalistic constraint we placed on our theory of content. In particular, she
distinguishes $clouds_1$ from $clouds_2$, yet acknowledges that they are “of the same physical type”—yet our naturalistic constraint restricts our attention to naturalistically defined categories, such as physical types! More importantly, however, Millikan’s account fails because it is simply irrelevant. In order to understand how humans (and animals) use clouds as a sign of rain, we need to understand the relationship between storm-clouds ($= clouds_1 \cup clouds_2$) and rain, not between $clouds_1$ and rain; people and animals can’t detect $clouds_1$, they can only detect $clouds_1 \cup clouds_2$. As such, Millikan’s account is only relevant insofar as it helps us to understand this relationship, and here her analysis says nothing more than storm-clouds and rain are statistically correlated within the relevant spatiotemporal region.

At this point, it looks as if Millikan can’t actually explain natural signs except in terms of statistical correlation, which she rejects. Dretske, by trying to get something stronger (signified with probability 1 given signifier), simply defines a situation which never obtains. Neither has provided a definition of natural signs which explains any of the commonly accepted examples of natural signification (including those they themselves discuss).

I think the difficulties in both views can be cleared up by acknowledging a distinction from the literature on stochastic causality.

The probabilistic treatment of causality leads immediately to a distinction between causal talk referring to population variables, or “property causality”, and causality between single events, often called “token” or “aleatory” causality. (Galavotti, 2001, 2)

“Property causality” may also be thought of as causal relations between types.

[Probabilistic theories of causality] are generally intended as theories of so-called type causation: that is they are intended to capture causal claims that relate types of events or properties such as “impacts of rocks cause windows to break” and “smoking causes lung cancer”. (Woodward, 2001, 39)

The crucial point for understanding how Grice’s discussion has led Dretske and Millikan astray is this: Grice discusses token instances of natural meaning, but the relation of natural meaning in general is one which must hold between types.
Once we have the distinction between type and token causality on the table, we can see that Grice’s analysis doesn’t necessarily apply to a general theory of natural signs. In fact, let’s investigate some Grice-style sentences, involving natural meaning between types.

3. This kind of spots usually means measles, but since your son has been exposed to Rubella, I’d like to run some more tests.

4. Dark, low clouds like this mean rain, but I don’t think we’ll see any rain today.

Once we move from looking at token cases of natural meaning to type cases, the factivity of “meaning” becomes more questionable. At the very least, we easily speak of one type of event “usually” or “ordinarily” meaning another, and this implies that the relationship between the two is ultimately stochastic.

The rest of this chapter will develop a theory of natural content based on statistical correlation between event types. This analysis applies in a straightforward manner to the classic examples of natural signs. Furthermore, I believe the above discussion demonstrates that it is not inconsistent with Grice’s argument. I myself am unconvinced that Grice intended his analysis to apply to natural signs rather than just uses of the term “meaning” in natural language. However, whatever his intent, the crucial features of natural meaning he discovered do not seem to extend to the type case.

Before presenting the positive theory, it may be worthwhile to briefly discuss why both Dretske and Millikan insist that “mere correlation” isn’t enough to get signification. For Dretske

The power of signs to mean or indicate something derives from the way they are related to what they indicate or mean. The red spots all over Tommy’s face mean that he has the measles, not simply because he has the measles, but because people without the measles don’t have spots of that kind. In most cases the underlying relations are causal or lawful in character. There is, then, a lawful dependency between the indicator and the indicated, a dependency that we normally express by conditionals in the subjunctive mood: if Tommy didn’t have the measles, he wouldn’t have those red spots all over his face. (Dretske, 1988, 56)
It is unfortunate that Dretske’s example is simply false (see Appendix A.2). The source of this falsity is not an idiosyncratic feature of the measles case, but holds of any standard or obvious example of natural signification. The crucial point here is that one can have a “lawful dependency” between two event types without the probability of one occurring given the other being 1. In the case of smoke and fire, for example, both are caused by the decomposition of solids when heated (Appendix A.1). If the conditions are right, smoke can occur without fire and vice versa, but the correlation between the two is very high because they are produced by a common cause.

In the case of measles and spots, we don’t know the exact causal relationship, just that there is a high correlation between the two. This is in fact the general case, and the motivation behind stochastic causality: the vast majority of causal claims in the social sciences, economics, psychology, biology, anthropology, etc. are claims about two correlated event-types between which the precise mechanism of causal influence is unknown. One of the most famous examples here is the claim that “smoking causes cancer”. This claim is so firmly believed that it is used to justify taxes and laws which restrict smoking behavior. (I have even seen it used to justify accusations of “irrationality” by academics against their smoking peers.) But the simple fact of the matter is: some people smoke and never acquire cancer, other people never smoke yet die of lung cancer. So, we need a stochastic notion of causality (and “lawlike dependence”) at the property or type level in order to preserve standard causal discourse. One is welcome to maintain a deterministic theory of causality at the token level if one wishes, but a theory of natural information must consider the relationship between types.

Millikan rejects correlation as an account of natural meaning for an entirely different reason. As we saw above, she acknowledges that correlation is the correct analysis for understanding how organisms learn from and track natural signs. Millikan’s worry is that correlation cannot be defined in a non-arbitrary fashion:

The notion of correlational information is empty unless a reference class for the correlation is specified, and there seems to be no way to specify such a reference class except arbitrarily. Correlations exist or fail to exist depending upon the reference class one chooses. If no single natural or
Millikan is absolutely right that correlation is only meaningful with respect to some sample set (this is how I interpret her “reference class”). For example, if we look only at the set \{George Burns\}, we will find zero correlation between smoking and lung cancer. If we wish correlations to depend upon relative frequencies of event occurrence in nature, then sample sets can be defined by spatiotemporal regions. Millikan’s point is that gerrymandered spatiotemporal regions will produce radically different assessments of natural signification.

However, I think Millikan is wrong in thinking that just because correlation is only well-defined with respect to a sample set (or spatiotemporal region) that it cannot provide an adequate foundation for a theory of natural information. Spatiotemporal regions are the kind of thing that simply exist, as are the relative frequencies of events within them. Sure, there may be no “natural” constraints on choosing one region or another, but we don’t need to. Natural signs are significational relationships out there in the world waiting to be used, just as spatiotemporal regions are out there waiting to be identified and distinguished. If an organism comes along and begins to track a natural sign relationship in the world, then the spatiotemporal region within which correlation must hold is well-defined: it is just that region which the organism visits throughout his lifespan.

So, correlation is our best candidate for characterizing the relationship of natural meaning. However, how can we use facts about the relative frequencies of events occurring within some region to define the informational content of a specific event? This is the subject of the remainder of this chapter.

### 5.3 Basic Concepts of Information Theory

Before we can define informational content, it will help to have a brief overview of the basic concepts in information theory. We will begin with the definition of a probability distribution, then we’ll examine how information is defined from probability. Next we
will look at two important measures of quantity of information, *entropy* and *cross-entropy*. We will conclude with a brief discussion of the role of these measures in inference, in particular, the axiomatically derived inferential mandate to minimize cross-entropy.

### 5.3.1 From Probability to Information

The public history of information theory began in 1948 with Claude E. Shannon’s seminal paper “The Mathematical Theory of Communication.” Shannon investigates the mathematical properties of a communication channel (such as a phone line or telegraph wire). Beginning with an analysis of a discrete channel, Shannon next considers the properties of information in a noisy channel, and finally develops a theory of continuous information. Shannon introduces the fundamental notion of *entropy*, to be discussed in the next section. (Dretske’s presentation of information theory in the first two chapters of Dretske (1981) summarizes the basic ideas in the first two chapters of Shannon and Weaver (1949), though using some non-standard terminology (see e.g. Good, 1983).)

There is also a “secret” history of information theory, however. Alan Turing and his team at Bletchley Park developed many of the key notions of information theory during the Second World War in their work decoding the signals of the Enigma machine. Many of the techniques developed during this period are discussed by I. J. Good (1979, but see also Good, 1950 and Good, 1965). Unfortunately, however, the full record of those endeavors remains classified.

The basic idea behind Shannon’s approach and that at Bletchley Park is exactly the same. We start with a background probability distribution. If we are attempting to decode Enigma signals, this may be a probability distribution over letters or words in German, for example one derived from the relative frequencies of letters and words in a standard corpus; if we are trying to analyze the channel conditions in American phone lines, it may be a probability distribution over letters or words in English. However, the basic approach doesn’t depend in any way on language, it merely assumes an antecedent probability distribution.
Next, we make an observation, this could be a signal received across a telegraph wire or an intercepted Enigma message. We want to have a measure of the information in our observation in order to quantify the evidence it gives us for certain inferences. In the most basic case, we may want to form an inference about whether the signal arrived is the same as the signal sent, or if it has been corrupted by noise in the communication channel. In the Bletchley Park case, we wish to infer the Enigma setting used (as this will allow us to decode all intercepted messages for the day).

In order to formalize these ideas, we begin with some finite set \( \Omega \) and an algebra \( \mathcal{A} \) over that set.

**Definition 5.3.1** Given a set \( \Omega \), \( \mathcal{A} \) is an *algebra* of sets on \( \Omega \) if it is nonempty and for every \( A, B \in \mathcal{A} \)

1. \(-A \in \mathcal{A}\)
2. \(A \cup B \in \mathcal{A}\)

Essentially, an algebra over \( \Omega \) is just a family of subsets of \( \Omega \) closed under union and complementation. (If we allow \( \Omega \) to be infinite, we must put more constraints on the type of algebra we allow; I will suppress these details here in the interests of clarity.) Notice that it is an immediate consequence of this definition that \( \Omega \in \mathcal{A} \). \( \mathcal{A} \) is nonempty, so there exists some \( A \in \mathcal{A} \), but this implies (by 1) that \(-A \in \mathcal{A} \). By 2, \( A \cup -A = \Omega \in \mathcal{A} \).

**Definition 5.3.2** Given a set \( \Omega \) and an algebra over that set \( \mathcal{A} \), a function \( P : \mathcal{A} \to [0, 1] \) is a *probability measure* if for all \( A, B \in \mathcal{A} \)

1. \( P(A) \geq 0 \)
2. \( P(\Omega) = 1 \)
3. if \( A \cap B = \emptyset \), then \( P(A \cup B) = P(A) + P(B) \)

All the standard properties of probability (Bayes' rule, etc.) are consequences of this definition. Finally, we can define a probability space (again, for the sake of simplicity, restricting ourselves to the finite case):
Definition 5.3.3 \( \langle \Omega, \mathcal{A}, P \rangle \) is a finite probability space if

1. \( \Omega \) is a finite set
2. \( \mathcal{A} \) is an algebra over \( \Omega \), and
3. \( P \) is a probability measure over \( \mathcal{A} \)

As discussed in Chapter 1, a random variable is just a function defined on a probability space. Random variables provide a convenient way to model observations. For example, if we are trying to decode Enigma messages it might be convenient to take \( \Omega \) to be the set of all possible finite sequences of roman characters, this is the space of all possible messages we might intercept. The act of intercepting a message and checking it for some feature of interest can be modeled with a random variable \( X \). For example, suppose we are interested in all those messages which begin with the same three characters, e.g. “the”. The messages which begin with “the” define a subset \( M \subset \Omega \), and we might wish to define the random variable \( X \) such that for any message \( m \), \( X(m) = 1 \) if \( m \in M \) and 0 otherwise.

Notice that the probability distribution over the outcomes of a random variable is defined in terms of the probability distribution over \( \Omega \). If \( p \) is the probability distribution over \( X \), then it is easy to see that \( p(X = 1) = P(M) \) and \( p(X = 0) = P(\neg M) \). Sometimes it will be helpful to look at the probability distribution over a random variable rather than the variable itself; we notate the probability distribution over \( X \) by \( X \sim p \).

Now suppose we have a finite probability space \( \langle \Omega, \mathcal{A}, P \rangle \) and we make an observation \( A \in \mathcal{A} \). We are interested in what the observation of \( A \) tells us about other possible events in \( \mathcal{A} \). This is the concept we will define in Section 5.4, the informational content of \( A \). We can begin here by asking a much simpler question: what is the quantity of information in an observation of \( A \)?

Intuitively, the quantity of information in \( A \) should be a function of its probability.\(^1\) Very low probability events give us large amounts of information, whereas very high probability events give us small amounts of information. In the limit, we learn nothing

\(^1\)The discussion in the next two paragraphs closely follows that of Osteyee and Good, 1974, 8–9.
if the probability of an event is 1 and a seemingly infinite amount if the probability
of an event is 0. I learn little when I observe that the sun has risen; I learn a great
deal when I observe a midget riding a motorcycle through my living room. These
considerations inspire the definition of the function $I$, the quantity of information in
an event.

**Definition 5.3.4** Given a finite probability space $\langle \Omega, \mathcal{A}, P \rangle$, the *quantity of information in* $A \in \mathcal{A}$ is given by a function $I : \mathcal{A} \to \mathbb{R}^+$ defined as

$$I(A) = -\log P(A)$$

This function satisfies our rough intuitions, i.e. $I(\Omega) = 0$, $I(A) \geq 0$ for all $A \in \mathcal{A}$,
and if $P(A) = 0$, $I(A) = \infty$.

This definition of the quantity of information in an event has another very desirable
property. If two events $A$ and $B$ are *independent*, then we learn nothing about
the occurrence of one having observed the other (i.e. if $A$ and $B$ are independent,
$P(A|B) = P(A)$ and vice versa).

**Definition 5.3.5** Given a finite probability space $\langle \Omega, \mathcal{A}, P \rangle$, events $A_1, \ldots, A_n \in \mathcal{A}$ are *independent* if

$$P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i)$$

We would like for the function $I$ to satisfy the intuition that information from the
observation of independent events should sum. Since the occurrence of $A$ doesn’t tell
us anything about the occurrence of $B$ (and vice versa), if we observe both $A$ and
$B$, we should learn everything we would have from $A$ alone plus everything we would
have from $B$ alone. In fact, this is the reason behind taking the logarithm as the log
function has the property that $\log(A \times B) = \log A + \log B$. The following observation
is an immediate consequence of this fact.

**Observation 5.3.6** If $\langle \Omega, \mathcal{A}, P \rangle$ is a finite probability space and $A_1, \ldots, A_n \in \mathcal{A}$ are
independent, then
\[
I(\bigcap_{i=1}^{n} A_i) = \sum_{i=1}^{n} I(A_i)
\]

It is important to notice here that all these desirable properties hold no matter what basis we use for the logarithm. Various figures have used $e$ or 10 in the past, but the most popular basis in the contemporary presentation of information theory is 2 due to the important role of binary alphabets in computer science. However, the choice of basis is essentially just a choice of unit, and it can be shown that any information function which satisfies Observation 5.3.6 and the desideratum that $P(A) < P(B)$ implies $I(A) > I(B)$ is proportional to $-\log P(A)$ (Osteyee and Good, 1974, 9).

Good attributes the first use of a logarithmic scale to measure information to Turing, who got the idea from an analogy with the decibel scale of loudness (Good, 1979). Good himself first published this analysis of information in 1950 (see esp. Chapter 6). The basic idea was developed simultaneously and independently in the United States by Shannon (Shannon and Weaver, 1949).

5.3.2 From Entropy to Inference

The function $I$ gives us a measure of the quantity of information in an event. By itself, this doesn’t do much but measure how much we learned from an event. What we’d really like to know is how we should change our expectations / beliefs / actions on the basis of our observation. The first important concept for achieving this end is entropy.

Shannon and Weaver (1949) model a communication channel as a Markov chain. A Markov chain is a sequence of time-indexed random variables such that $X_{t+1}$ depends only upon $X_t$. Many systems in nature can be thought of as Markov chains. Consider, for example, molecules of gas trapped in a box. At any time $t + 1$ the positions and velocities of all the particles depend only upon their positions and velocities at time $t$. In fact, if we take a probabilistic perspective on causality, many chains of causal influence are naturally modeled as Markov chains (for some detailed examples
see Good, 1961 or Suppes, 2001).

If we wish to model a communication channel as a Markov chain, then at each step in time, the channel produces some symbol (let's say a character in the roman alphabet) with a probability dependent only upon the previous character. If the language being sent along the channel is English, for example, we would expect upon seeing “q” at time $t$ to see “u” at $t + 1$ with a very high degree of probability. If a Markov chain depends upon the previous $n$ steps in its history, then it is an $n$ order Markov chain. Shannon and Weaver discuss how the appearance of English text can be produced with relatively low order Markov chains (39–48). It is surprising, for example, just how “English-like” text generated by a third order Markov chain over letters can be. So, even though the message coming through a communication channel in general is not actually produced by a Markov chain, taking it to be so allows one to develop a qualitatively realistic model of the distribution of symbols one encounters coming from a real source.

Since Shannon considers the sequence of symbols he observes as being produced by a Markov chain, he can take each one to be the outcome of a “choice” determined by the probability distribution associated with each random variable in the chain. He wants a measure of how much “choice” the system has, or, equivalently, how much uncertainty we have about the symbols generated. This measure, $H$, should depend solely on the distribution of probabilities at the source. If $p_i$ is the probability of event $i$ occurring and there are $n$ possible events, then Shannon identifies three conditions to place on $H$:

1. $H$ should be continuous in the $p_i$.
2. If all the $p_i$ are equal, $p_i = \frac{1}{n}$, then $H$ should be a monotonic increasing function of $n$. With equally likely events there is more choice, or uncertainty, when there are more possible events.
3. If a choice be broken down into two successive choices, the original $H$ should be a the weighted sum of the individual values of $H$... (Shannon and Weaver, 1949, 49)

Condition 1 simply says that tiny changes in probabilities should produce only tiny changes in $H$. 2 is self explanatory. In order to explain 3, Shannon uses the example
in Figure 5.1. The essential idea is that a probability distribution should be assigned the same $H$ whether it was reached in one step, or in multiple steps. In Figure 5.1, the process on the left generates a distribution in a single step, the process on the right generates the same distribution in two steps. Consequently, condition 3 requires that $H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(\frac{2}{3}, \frac{1}{3})$.

Shannon proved the following

**Theorem 5.3.7 (Shannon)** The only $H$ satisfying the three above assumptions is of the form

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

where $K$ is a positive constant. (49–50)

$H$ is called the entropy of the information source, and since $K$ is merely a scaling factor, it is usually dropped.

If I am playing a game of chance, my expectation about what I will win is the sum of possible winnings weighted by the probability that each will obtain. If I flip a fair coin, to win $10 if heads and to lose $5 if tails, my expected winnings are $\frac{1}{2} \cdot 10 + \frac{1}{2} \cdot (-5) = 2.50$. Given some quantity $x$ and a probability distribution $p$, we denote the expectation of $x$ given $p$ by $E_p(x)$. Entropy is expected information.

**Observation 5.3.8** Given a finite probability space $\langle \Omega, \mathcal{A}, P \rangle$, and a random variable over that space $X \sim p$ which takes values $x_1, \ldots, x_n$,

$$H(X) = E_p I(x_i)$$
The entropy of a probability distribution is at its maximum when the probability of all event is equal. If $X \sim p$ is a random variable that can take $n$ distinct values $x_1, \ldots, x_n$, then $H(X)$ is maximum when $p(x_i) = \frac{1}{n}$ for all $0 \leq i \leq n$.

How can entropy help us form inferences? In 1957, E. T. Jaynes proposed the principle of maximizing entropy. If we don’t know the probability distribution generating signals from an information source, then we should infer the distribution which maximizes entropy consistent with our observations. Jaynes points out that this principle is a natural extension of Laplace’s principle of insufficient reason. Laplace argued that in the absence of any information about a system, we should take the probabilities of all events to be equal, i.e. if there are $n$ possible outcomes, we should take $p(x_i) = \frac{1}{n}$ for all $0 \leq i \leq n$. Jaynes discusses:

The principle of maximum entropy may be regarded as an extension of the principle of insufficient reason (to which it reduces in case no information is given except enumeration of the possibilities $x_i$), with the following essential difference. The maximum-entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally noncommittal with regard to missing information, instead of the negative one that there was no reason to think otherwise. Thus the concept of entropy supplies the missing criterion of choice which Laplace needed to remove the apparent arbitrariness of the principle of insufficient reason, and in addition it shows precisely how this principle is to be modified in case there are reasons for “thinking otherwise.” (Jaynes, 1957, 623)

Jaynes uses the principle of maximum entropy to provide a subjectivist foundation for entropy assessments in statistical mechanics. The essential point for him is that the principle of maximizing entropy allows for inferences that do not import any bias. The principle of maximizing entropy “is the least biased estimate possible on the given information; i.e. it is maximally noncommittal with regard to missing information” (Jaynes, 1957, 620).

Jaynes’ principle actually turns out to have a more general form based on a generalization of the notion of entropy. Entropy is a property of an information source which indicates our degree of uncertainty about the signals we might observe. Suppose we already have some sense of the probability distribution which is at the source.
When we observe a set of signals, then, we are not so much interested in their entropy as in their entropy relative to our expectations about the distribution at the source. We can think of this as the extent to which the “actual” probability distribution at the source (say, \( p \)) diverges from our assumed distribution (say, \( q \)), which we notate \( D(p||q) \).

**Definition 5.3.9** Given a finite probability space \( \langle \Omega, A, P \rangle \) and random variables over that space \( X \sim p \) and \( X' \sim q \) which can take \( n \) values \( x_1, \ldots, x_n \), where \( \mu_i \) indicates the probability of value \( x_i \) for \( \mu \in \{p, q\} \) and \( 0 \leq i \leq n \), the divergence of \( p \) from \( q \) is

\[
D(p||q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}
\]

This measure of the “distance” between two probability distributions is usually attributed to Kullback and Leibler (1951), though a similar measure can be found in Good (1950), 75. It is sometimes called the Kullback-Leibler divergence, the information for discriminating in favor of \( p \) against \( q \), relative entropy, cross-entropy, or expected weight of evidence.

**Observation 5.3.10** \( D \) is not symmetric, i.e., in general, \( D(p||q) \neq D(q||p) \)

Kullback (1959) first states the principle of minimizing cross-entropy as a guide for inference. In the special case where \( q \) is the distribution with maximal entropy, i.e. \( q_i = \frac{1}{n} \) for all \( 0 \leq i \leq n \), the principle of minimizing cross-entropy is equivalent to the principle of maximizing entropy.

**Theorem 5.3.11** Given a finite probability space \( \langle \Omega, A, P \rangle \), a random variable \( X \sim q \) which may take \( n \) values \( x_1, \ldots, x_n \) such that \( q(x_i) = \frac{1}{n} \) for all \( 0 \leq i \leq n \), and some constraints \( C \) on a probability distribution \( p \) over \( x_1, \ldots, x_n \) such that \( p \neq q \), the unique \( p \) which satisfies \( C \) and minimizes \( D(p||q) \) and that which satisfies \( C \) and maximizes \( H(p) \) are identical.

**Proof.** Maximizing \( H(p) = -\sum p_i \log p_i \) is equivalent to minimizing \( \sum p_i \log p_i \). This in turn is unaffected by the addition of an arbitrary constant, so is equivalent to minimizing \( \sum p_i \log p_i + c \). If \( q \) is the maximal entropy distribution, then \( D(p||q) \)
is of the form $\sum p_i \log p_i + c$ since

$$D(p||q) = \sum p_i \log \frac{p_i}{q_i} = \sum p_i \log p_i - \sum p_i \log q_i = \sum p_i \log p_i - \sum p_i \log \frac{1}{n} = \sum p_i \log p_i - \log \frac{1}{n} \quad \text{and for fixed } n, -\log \frac{1}{n} \text{ is a constant.}$$

Theorem 5.3.11 is proved more rigorously by Shore and Johnson (1980), who also provide an axiomatic justification for the principle of minimizing cross-entropy. Shore and Johnson prove that any inference principle which satisfies four intuitively plausible constraints must produce results equivalent to those produced by the principle of minimizing cross-entropy. Here is their informal characterization of the four constraints:

1. *Uniqueness*: The result should be unique.
2. *Invariance*: The choice of coordinate system should not matter.
3. *System Independence*: It should not matter whether one accounts for independent information about independent systems separately in terms of different densities or together in terms of a joint density.
4. *Subset Independence*: It should not matter whether one treats an independent subset of system states in terms of a separate conditional density or in terms of the full system density. (Shore and Johnson, 1980, 27)

Essentially, the point of these constraints is to ensure that no artifacts of our choice of notation, units, or order of update will infect our inferential procedure. Shore and Johnson finally put the intuitive justifications of Jaynes, Kullback, and many others into an axiomatic framework. In the next section, we'll see how the concepts developed here can be used to define informational content.

### 5.4 Introducing the $s$-vector

In this section I will finally provide an explicit definition of informational content derived from information theory. I call the formal object which characterizes the informational content of an event an *$s$-vector*. This formal object is derived from the inferential principles of Shannon information theory; it has been suggested recently by Brian Skyrms (as discussed in Section 5.4.1), and it is independently motivated by empirical work on word co-occurrence statistics by Bullinaria and Levy (2007), who
call it a “semantic vector” (see Section 5.4.2). So, the “s” in s-vector may be taken to stand for “Shannon”, “Skyrms”, or “semantic” as your fancy chooses.

Philosophical lore states that information theory does not provide an account of informational content, but only of informational quantity. This view derives in part from the attitude of Shannon himself, who states that the engineering problem of communication theory has nothing to do with content or meaning.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design. (Shannon and Weaver, 1949, 31)

As we saw in the previous section, however, information theory has proved to have uses far beyond its telecommunication roots. And furthermore, during the “secret” history of information theory at Bletchley Park, it was already being used to drive inferences. The purpose of these inferences was ultimately to determine the content of intercepted Enigma messages, so certainly information theory can be used in situations where “semantic aspects” are not, as Shannon claims, “irrelevant”.

Motivated by Shannon’s claims and Grice’s constraint, Dretske developed his own “semantic theory of information”. The essential idea depends upon Dretske’s interpretation of Shannon’s theory; an early passage is telling here:

Information theory identifies the amount of information associated with, or generated by, the occurrence of an event ... with the reduction in uncertainty, the elimination of possibilities, represented by that event or state of affairs. (Dretske, 1981, 4, emphasis added)

Right on page 4, we see Dretske rejecting the most essential feature of information theory in a move that has been taken to demonstrate deep insight into the nature of
information (see for example Barwise and Perry, 1999; van Benthen and Martinez, 2008; or the work of Luciano Floridi, e.g. 2004). Dretske glosses “the reduction of uncertainty” as “the elimination of possibilities”, but this is problematic as an inference principle since it demands a separate procedure for reinstating possibilities should one ever be eliminated in error. The beauty of probability theory is that it allows for changes in one’s degree of certainty without elimination of possibilities. Of course, a probabilistic inference procedure like conditionalization can eliminate possibilities by assigning them probability zero, but it can also make more fine-grained adjustments. The advantage of assigning a very low probability to a possibility (rather than simply zero) is that one can later use the same inference procedure to increase uncertainty by raising its probability again. Since they are based in probability theory, information theory and minimization of cross-entropy also enjoy this feature. This is not to say that Dretske’s analysis is not of independent interest (and it has proved its worth in situation semantics and dynamic logic), but it involves a much coarser notion of uncertainty than that found in the mathematical theory of information.

Dretske’s analysis of “reduction of uncertainty” (again, plus his adherence to Grice’s constraint) motivates his identification of the informational content of a signal with propositions which hold with probability one conditional on that signal.

Informational content: A signal $r$ carries the information that $s$ is $F = \text{The conditional probability of } s \text{'s being } F, \text{ given } r \text{ (and } k \text{), is 1 (but, given } k \text{ alone, less than 1)}^2$ (Dretske, 1981, 65)

Dretske’s stipulation that informational content demands that posterior probability be one has been criticized as implausible (Arbib, 1983; Armstrong, 1983; Ginet, 1983; Sayre, 1983; Sosa, 1983), uninteresting (Kyburg, Jr., 1983; Levi, 1983), and counterintuitive (Rozeboom, 1983; Suppes, 1983). The most important concern for our purposes is this: Dretske’s analysis of informational content applies to nothing of interest in the world. It cannot capture our best intuitive examples of natural signs (Appendix A); it cannot explain the informational relationship between perception and world (which is inherently stochastic as discussed in Chapter 2); and it cannot support safe reasoning about the world.

$^2k$ here is background knowledge.
Given our analysis in Section 5.2, Grice’s considerations do not constitute a legitimate restriction on a theory of natural meaning. In the remainder of this section we will define and defend an analysis of informational content in terms of $s$-vectors which i) is sensitive to the details of information theory; ii) applies directly to all intuitive examples of natural information, and iii) can serve as the basis for safe inferential procedures. Finally, in Subsection 5.4.2, we will see that this analysis has received empirical support from research on word co-occurrence statistics.

### 5.4.1 Theoretical Motivation

Skyrms (2010) exemplifies the perspective taken here. Just like the author, Skyrms begins with the intuition that Dretske has given up on information theory too soon. He expresses many of the same sentiments we have seen above:

Some philosophers have looked at information theory and have seen only an answer to the question of quantity. They do not see an answer to the question of content—or, to use a dangerous word, meaning—of a signal. As a result they move to a semantic notion of information, where the informational content in a signal is conceived as a proposition. The information in a signal is to be expressible as “the proposition that________.” Signals then, in and out of equilibrium, are thought of as the sorts of things that are either true or false. Dretske takes that road and, as he himself says, it reduces the role of information theory to that of a suggestive metaphor. Others have followed his lead. I believe that we can do better by using a more general concept of informational content. (Skyrms, 2010, 34)

Skyrms examines the emergence of signaling systems from initially random interactions amongst agents with shared goals. It turns out the signaling systems emerge in models using very simple reinforcement learning to model changes in signaling behavior. This project is naturalistic in the same sense as the present work—Skyrms explains the existence of signaling games by appealing entirely to simple descriptive models of agent interaction over long time periods. Given such a signaling system, can information theory provide an account of the content of its signals? Skyrms’ essential insight is that
The *informational content* of a signal consists in how the signal affects probabilities. (34)

Skyrms thinks that Kullback-Leibler divergence (cross-entropy) holds the key to characterizing both informational quantity and informational content. Suppose you and I are part of an evolved signaling game. By assumption, there is some stable probability distribution over states of interest in the world, call it $p$. When I receive a signal from you, call it $\sigma$, then I learn something about the world. In particular, my assessment of the probability distribution over states of the world changes. Call this new distribution $p_\sigma$. The Kullback-Leibler divergence tells me just how much my new probability distribution differs from the original distribution. If there are $n$ states of interest in the world, $x_1, \ldots, x_n$, then the divergence of $p_\sigma$ from $p$ is just:

$$
\sum_{i=1}^{n} p_\sigma(x_i) \log \frac{p_\sigma(x_i)}{p(x_i)}
$$

This is the *quantity* of information in the signal.

The Kullback-Leibler divergence is a weighted average over the log ratio of probability assignments for each possible event. *If* the log ratio of probabilities is an appropriate measure of the quantity of informational change about the event $x_i$ in the signal $\sigma$ (an issue to be discussed further below), then the set of all such log probability ratios characterizes *everything the signal tells us about the world*. This complete characterization of informational content is what Skyrms calls “how the signal affects probabilities”. If we buy this we can see that the Kullback-Leibler divergence measures quantity of information by averaging over the distinct pieces of information provided by the signal for each $x_i$, effectively erasing the details of each of these pieces of information in the process. In order to provide a complete characterization of the informational content of $\sigma$, then, we merely need to organize the log ratio probabilities into a single formal object which preserves them for each $x_i$. A natural choice here is a vector, since vectors organize distinct quantities into a single formal object. This is what I call the $s$-vector of $\sigma$.

$$
\text{s-vector}(\sigma) = \left\langle \log \frac{p_\sigma(x_1)}{p(x_1)}, \log \frac{p_\sigma(x_2)}{p(x_2)}, \ldots, \log \frac{p_\sigma(x_n)}{p(x_n)} \right\rangle
$$
Claim: $s$-vector($\sigma$) characterizes the informational content of $\sigma$.

The essential insight is this: Yes, information theory can only tell us about quantities of information, but if we know individually the quantity of information in a signal about each possible state, then this is all the information contained in that signal. So, the collection of all these quantities together tells us everything about the world which the signal tells us, i.e. its content. The $s$-vector formalizes this intuition.

Now, Skyrms is worried about naturally emergent signaling systems. Are natural signs anything like signaling systems? Is it legitimate to use Skyrms’ analysis of content as a theory of the meaning of natural signs? I think the answer is an emphatic yes. The original insight of Augustine was that natural signs are like a language of nature waiting to be read—signals waiting to be interpreted. More importantly, they exemplify all the features needed to define an $s$-vector. By assumption, there is some stable frequency of event-types in the world. If we witness an instance of an event-type, this changes the probabilities of other event-types away from this stable distribution. Suppose the probability of encountering fire within a region is $P(\text{fire}) = x$. If we encounter smoke, that probability that fire is present changes to $P_{\text{smoke}}(\text{fire}) = y > x$, consequently

$$\log \frac{P_{\text{smoke}}(\text{fire})}{P(\text{fire})} > 0$$

And, as expected, the $s$-vector tells us that smoke is a sign of fire and, furthermore, it tells us exactly good of a sign of fire smoke is.

Of course, it is one of the features of natural signs that they can signify many different things. “Spots”, more properly a particular quality of rash, for example, tell us about measles (they raise the probability of measles), but they also tell us about rubella, glandular fever, and drug allergies, i.e. the probability of all these states obtaining is raised by the presence of this type of rash (see Appendix A.2). Consequently, in the $s$-vector for this rash type, the values associated with measles, rubella, etc. will all be positive. If the rash appears in conjunction with another sign of measles, say Koplik’s spots, then the $s$-vector for the two together will spike in value at a single state, measles. And this is a fair first approximation to how doctors
Chapter 5. The Informational Content of Natural Signs

Actually reason: they look at a confluence of symptoms and infer to the single cause the probability of which deviates most strongly from baseline frequencies.

Although he doesn’t develop the issue in detail, Skyrms himself acknowledges that his approach applies directly to the content of natural signs:

[N]either intentionality nor teleosemantics is required to give an adequate account of the informational content of signals. Here I stand with Dretske. The information is just there. At this point some philosophers will say “You might as well say that Smoke carries information about fire.” Well, doesn’t it? Don’t fossils carry information about past life forms? Doesn’t the cosmic background radiation carry information about the early stages of the universe. The world is full of information. (Skyrms, 2010, 43–4)

This leaves a final issue: is there an independent theoretical motivation for taking the log ratio of probabilities? What we want is a measure of relative information, i.e. information relative to stable background frequencies. The probability of the sun rising tomorrow is very high given that I ate cornflakes this morning, but my breakfast doesn’t give me information about tomorrow’s sunrise because the probability of the sun rising tomorrow was already high.

It would be nice to have the same kind of axiomatic justification for a measure of relative information as we saw exists for information $I$ and cross-entropy $D$. At the very least, the log ratio of probabilities exhibits some very nice qualitative features. Consider, for example, two events $e_1$ and $e_2$. The proposal is that the quantity of information about $e_2$ given by $e_1$ relative to a background distribution $p$ be given by

$$RI_p(e_2 : e_1) = \log \frac{p(e_2|e_1)}{p(e_2)}$$

This formula has several desirable properties. If $p(e_2|e_1) > p(e_2)$, $RI_p(e_2 : e_1)$ is positive; if $p(e_2|e_1) < p(e_2)$, $RI_p(e_2 : e_1)$ is negative, approaching $+\infty$ and $-\infty$ respectively as the distance between the two distributions becomes greater. If $p(e_2|e_1) = p(e_2)$, then $RI_p(e_2 : e_1) = 0$, capturing the intuition that there is no information about $e_2$ in $e_1$ if it doesn’t change the probability of $e_2$ at all.
Unlike $D$, $RI$ is symmetric; for this reason it is sometimes called the \textit{pointwise mutual information} between $e_1$ and $e_2$.\footnote{$3$} The symmetry of $RI$ is an immediate consequence of the fact that $p(A|B)p(B) = p(A \land B) = p(B|A)p(A)$:

$$RI_p(e_2 : e_1) = \log \frac{p(e_2|e_1)}{p(e_2)} = \log \frac{p(e_2 \land e_1)}{p(e_2)p(e_1)} = \log \frac{p(e_1|e_2)}{p(e_1)} = RI_p(e_1 : e_2)$$

This conforms well to the intuition that natural information is symmetric: if the presence of smoke tells me about the presence of fire, then the presence of fire tells me about the presence of smoke. The asymmetry in our usual attributions of natural signification comes from an asymmetry in our interests and abilities. I usually think of smoke as a sign of fire rather than vice versa because smoke is very easy for me to spot from a distance and fire is an event I care deeply about and wish to detect. If I am present at a fire, then of course it gives me information about the presence of smoke, but I don’t really care because what I care about is escaping the danger of the fire. Likewise with clouds and rain. I can easily spot clouds and learn about a potential future event of interest, rain. If it is already raining, then I don’t really care whether there are clouds present or not, although I do receive information about the presence of clouds. Anyone who has vacationed in north England’s “lake district” can confirm that, just as the presence of clouds does not raise the chance of rain to 1, the presence of rain does not raise the chance of clouds to 1 either.

So, $RI$ satisfies many of our intuitions about an appropriate measure of relative information. I have not been able to find the kind of axiomatic justification in the literature which exists for $I$ and $D$, however. I. J. Good has argued extensively in favor of a closely related measure, which he calls \textit{weight of evidence} (e.g. Good, 1950, Chapter 6, and Good, 1985). The weight of evidence in favor the hypothesis $H$
provided by the event $e$ is defined as

$$W_p(H : e) = \log \frac{p(e|H)}{p(e|\neg H)}$$

and it is the strong similarity between $RI$ and $W$ which inspires Good to call $D$ the expected weight of evidence.

However, there has recently been a great deal of subtle activity in the literature on confirmation theory making fine-grained distinctions between measures such as $RI$ and $W$. See, for example, Fitelson (1999), which demonstrates qualitative differences in the behavior of $RI$ and $W$ with respect to various desiderata which have emerged within confirmation theory. I suspect the growing body of literature on this topic will eventually result in an axiomatic justification of some particular measure of confirmation, potentially $RI$, but a consensus has yet to emerge on this matter.

Given the axiomatic justification for minimizing cross-entropy in Shore and Johnson (1980), it would be surprising if $RI$ cannot be justified in some intuitive sense. If $RI$ turns out to have some conceptually undesirable consequence and a different measure receives the support of the confirmation theory community, then this other measure must somehow be reconciled with the principle of minimizing cross-entropy. Nevertheless, as a theory of informational content, we need not rely solely on the intuitive justification of the $s$-vector provided by Skyrms, since his analysis receives empirical support from independent work on the use of relative frequencies in corpora to derive semantic models.

### 5.4.2 Empirical Support

Skyrms (2010) analyzes the emergence of signaling systems using various game theoretic communication models. However, his analysis of content plays no particular role in these models. Are there examples where an analysis of content along the lines of the $s$-vector has been used to perform tasks that we would expect a semantic theory to perform, i.e. is there any empirical justification for taking the $s$-vector as an analysis of content? As it turns out, the answer is yes, and support for the $s$-vector can be found in work developed completely independent of Skyrms’ analysis.
A literature has emerged recently on the use of word co-occurrence statistics to derive semantic representations. The basic idea is this: start with a large corpus and examine the contexts in which a given word appears within that corpus. There are a number of different ways to cash out context here; for example, one could consider the word immediately preceding the target word as its context. One would then count the relative frequencies for all target words of those words which immediately precede them and use these statistics as a semantic model. The success of this semantic model will depend upon various empirical tests. For example, we may be interested in whether this model can successfully distinguish interesting classes of words. We would expect nouns, for example, to be preceded by the word “the” much more frequently than verbs. Will distributional effects such as this be enough to correctly categorize words into nouns and verbs on the basis of our semantic model?

Those aspects of meaning most effectively addressed within this literature concern degrees of similarity or difference between words, e.g. categorization and synonymy tasks. Success on these tasks depends upon a measure of distance between the semantic representations of words. So there are two important questions here. First, how do we use relative frequencies in a corpus to define a semantic model of word meaning? Second, what is the appropriate distance metric on the space of meanings?

One of the most sophisticated and influential recent works in this literature is Bullinaria and Levy (2007). Unlike many other studies in this area, Bullinaria and Levy do not begin by deciding a priori on answers to these two questions. Instead, they consider a number of different answers to both questions and empirically determine the best combination of answers in terms of relative success on a set of semantic tasks. As it turns out, the semantic representation which worked best is the positive version of the s-vector as defined in the previous section. The remainder of this section will discuss the methods and results of Bullinaria and Levy (2007) in a little more detail.

All the semantic models considered by Bullinaria and Levy are vectors with components derived from relative word frequencies within the corpus. The context of interest is defined in terms of a window $W$ of words around the target word $t$. $n(c,t)$ is the total number of times a context word $c$ occurs within windows of size $W$ around
the target word $t$. The basic semantic vector for the word $t$ is just the vector of conditional probabilities defined by

$$p(c|t) = \frac{p(c \land t)}{p(t)} = \frac{n(c, t)}{\sum_c n(c, t)}$$

Of course we can also use this data to define the probability of a given word simpliciter by normalizing the counts with respect to window size and the total number of words in the corpus.

$$p(c) = \frac{1}{NW} \sum_t n(c, t)$$

(Bullinaria and Levy, 2007, 513)

The basic semantic vector summarizes the relative frequency of context words in the corpus around the target word. However, different representations of this same information are possible. Although they consider several, the most successful representation turned out to be the vector of positive pointwise mutual information (PMI). PMI was discussed above as the function $RI$. Positive PMI is the same as $RI$ except when $RI \leq 0$, in which case it is just set to zero; call this $RI^+$. So the most successful semantic vector found by Bullinaria and Levy is that proposed by Skyrms with all negative values set to 0.

Bullinaria and Levy consider a number of different distance measures between vectors, including euclidean, city block, cross-entropy, and many which are more esoteric. The most successful metric in combination with the vector of positive PMI turned out to be the cosine, defined as one minus the cosine of the angle between the two vectors. Effectively, the cosine measures the degree of similarity between the directions of two vectors. So, the most effective measure of distance between two words $t_1$, $t_2$ turned out to be
CHAPTER 5. THE INFORMATIONAL CONTENT OF NATURAL SIGNS

The combination of positive s-vector and cosine distance beat out all other combinations of representation and metric on three of the four tests employed by Bullinaria and Levy (and ranked very close to the top on the fourth). The four tasks employed by Bullinaria and Levy to measure the efficacy of their semantic representations were:

1. Multiple choice synonymy questions from TOEFL (Test of English as a Foreign Language).

2. Multiple choice distance comparison between related words; “typical related words are brother and sister, black and white, lettuce and cabbage, bind and tie, competence and ability.”

3. Semantic categorization: “Ten words were taken from each of 53 semantic categories (e.g., metals, fruits, weapons, sports, colors) based on human category norms [derived from results in experimental psychology], and the percentage of the 530 words that fell closer to their own category center rather than another was computed.”

4. Syntactic categorization: “The degree to which word vectors are closer to their own syntactic category center rather than other category centers is measured.”

Positive s-vector plus cosine distance achieved above 95% success on tasks 2 and 4, 85% success on task 1, and 80% success on task 3. It performed better than all other
combinations on tasks 1 through 3, and it performed better than previous attempts in the literature on those tasks which others had attempted (e.g. TOEFL, for which previous published success rates were below 75%, see Bullinaria and Levy, 2007, 515).

[Bullinaria and Levy also varied the size of $W$ and discovered that a window of size 1 on either side of the target word worked best (Bullinaria and Levy, 2007, 516–8).]

Of course, there are many differences between words and natural signs, and we would expect their semantic treatment to differ dramatically. However, the types of worry that naturally arise here are precisely those which involve qualitative differences between natural meaning and conventional meaning. Humans may make mistakes in their use of a term, but nature does not “make mistakes” with causal effects; humans may lie or deliberately abuse words for creative purposes, but nature can do neither of these with natural signs; etc. The models of Bullinaria and Levy are blind to such “errors” in the corpus, instead, they depend upon only a single feature of the corpus, the same feature available in the case of natural signs—“constant conjunction” in Hume’s elegant formulation. Correspondingly, those aspects of meaning which they succeed in capturing with these models are the very aspects we may expect to find holding for natural signs.

Just as we assume some causal story underlies the co-occurrence of events we observe in nature (smoke and fire, clouds and rain), some causal story does underlie the arrangement of words in a corpus. What is telling here is the success in distinguishing relevant consequences of this underlying causal story by merely looking at the co-occurrence of words. Of course, these consequences are all cashed out in terms of a single semantic relationship, degree of similarity. But this is the only semantic relationship we might reasonably expect to apply to natural signs. For example, a certain type of vibration in the ground and a certain type of noise in the underbrush may both be signs of the presence of an elephant. Of course vibrations and noises are themselves qualitatively very different types of entities, but the $s$-vector for this vibration type and that for this sound type will fall very close together by any reasonable distance metric. This is expected because both vibration and sound will deliver information about many of the same events. The work of Bullinaria and Levy demonstrates that this expectation is fulfilled by the $s$-vector.
So, the s-vector, which was already motivated by a priori considerations, receives empirical support from Bullinaria and Levy (2007) because they demonstrate it indeed has the features our analysis expects (e.g. effective similarity discrimination) for a data set which is qualitatively similar to natural events in the relevant respect (namely the points are distributed in accordance with an (unknown) causal pattern).

5.5 Conclusion

In this chapter, we saw that the content of natural signs can be derived from their relative frequencies of co-occurrence within a specified spatiotemporal region. The insights of Grice on natural meaning do not apply here because he analyzed a meaning relation between event tokens, yet natural signification is a relation between event types. The appropriate formal object for capturing the meaning of a natural sign is an s-vector. The s-vector finds philosophical support from inferential procedures in information theory and suggestive arguments by Brian Skyrms. It finds empirical support in the work of Bullinaria and Levy.

In Chapter 3 we saw that perceptual experience can be treated as natural measurement. However, just as natural signs can carry information about many distinct states of affairs, natural measurements can measure many distinct domains. This information comes in two distinct flavors. First, natural signs can carry information about mutually exclusive alternatives, each of which is positively correlated with the signifier, e.g. spots are a sign of measles, but also of rubella, glandular fever, etc. Second, natural signs may carry information about different stages (or correlates of stages) in their causal past. Suppose, for example, that in a jungle without any human inhabitants, the most common cause of fires is lightning. Then in that jungle, smoke is a sign of fire and it is a sign of lightning, not because the two are mutually exclusive, but because they fall along a possible causal chain ending in smoke.

Natural measurements, and, in particular, perceptual experiences, also carry both types of information. Suppose two distinct surface properties produce a qualitatively similar perceptual experience e. Then s-vector(e) will contain positive values associated with both surface properties. It will also assign positive values to different
stages in the causal path leading to that experience, e.g. to possible properties of the illuminant. Now that we have a characterization of the informational content of the experience, we need to find a way to single out one amongst the many states of affairs about which it contains information as the unique intentional content of the experience. This is the task we set ourselves in Chapter 1, and the principle of minimizing cross-entropy will be used to suggest a solution in the following chapter.
Chapter 6

Realism, Metamerism, and Beyond

6.1 Introduction

In our concluding chapter, we revisit the traditional color realist debate and propose a solution to the problem of deriving intentional content from informational content in the particular case of color experience.

The traditional question of color realism asks whether or not colors are properties of surfaces. I will argue that this question simply does not make sense within the framework for interpreting perceptual experience outlined in Chapter 3. I demonstrate an analogy between color perception and spatial perception; comparison with this less contentious domain will help make the point more transparent. Essentially, I will show that there is a sense in which “colors” are obviously “real”, a sense in which they obviously are not, and that there isn’t any interesting additional debate to be had about “realism” here since all parties in the debate agree to these two obvious claims.

However, there are issues which emerge in the color realism debate which are of serious interest. In particular, how efficient is color vision? How natural are the categories it identifies? These questions arise because of metamerism, the fact that physically distinct stimuli (“metamers”) can produce exactly the same perceptual experience. Yet at first blush, groups of metameric stimuli (whether they be spectral power distributions of light or surface reflectance properties) do not seem to form
natural categories. In Section 6.3, some strategies for addressing the problem of metamers are considered. Some of these strategies can be found in the philosophical literature, others emerge in rather specialized areas in color science which (to my knowledge) philosophers have yet to address.

Finally, we will extend Dretske’s notion of *primary representation* in accordance with the probabilistic theory of informational content introduced in the previous chapter. The essential proposal is that we can use cross-entropy as a measure of the efficiency of color perception. By looking for the stage in the causal chain from external world to internal experience which is most efficiently represented by color experience, we can identify the representational function, and thereby intentional content, of a color percept. Unfortunately, it is not possible in the present document to produce a definitive answer to the question of what color experience represents due to a lack of data and computational resources. However, the method introduced in Section 6.4 provides an *in principle* procedure for determining intentional content from purely descriptive facts about the world, which satisfies the naturalistic constraint motivating this project without appealing to teleology or evolution. In fact, the analysis presented here might be taken as independent evidence in favor of a particular selectionist hypothesis, contributing to the evolutionary analysis of function rather than simply parasitizing it.

This chapter will conclude with a summary and brief discussion of what has been achieved in the project as a whole.

### 6.2 The Irrelevance of Realism

There is a central concern shared by nearly all the papers collected here (and most contemporary philosophical discussions of color), namely, to answer the following questions: Are physical objects colored? And if so, what is the nature of the color properties? These questions form the problem of *color realism*. (Byrne and Hilbert, 1997, xi)

These words appear on the very first page of the introduction to the seminal collection *Readings on Color: The Philosophy of Color*. In this section, I will argue that *if*
we accept the claim that perception is measurement, then this question is at best irrelevant and at worst a distraction from legitimate philosophical and scientific issues.

One reason the question of color realism is taken to be so important is questions of veridicality: Does my experience represent the world as it is? I do not wish to deny the importance of this question, but rather to emphasize that it can in principle come apart from the question of whether or not physical objects are colored.

To see this, let’s return to the example of temperature. Suppose I measure an object and my thermometer reads 78° Fahrenheit. No one would think that the veridicality of this measurement turns upon the question of whether “objects really are numbered,” or whether “78” is a property of this particular object! Of course, we can casually ask whether the object really is 78° Fahrenheit, but this is a question about the accuracy of the measurement we performed, not the similarity between a point in the representing space (“78”) and the corresponding point in the represented space (the mean molecular motion of the target object).

As argued in Chapter 3, the case of color perception is exactly analogous. The crucial point to acknowledge here is this: in order to think clearly about measurement, we need to keep the measured domain and the measuring domain distinct. Just as we need to remember that numbers and temperatures are qualitatively different sorts of things in order to analyze whether our thermometric practice is consistent and well-founded, so also we need to acknowledge that external properties and our perceptual experience of them are distinct in order to better understand the relationship between the two.

Notice: we can measure a property with that exact same property, but even in this case, veridicality does not turn on the represented property being “the same” as the representing property. Consider the case of scale models. Suppose I have constructed an O gauge model train layout on which I’ve placed a scale model of the Eiffel Tower. Is my depiction of the height of the Eiffel Tower veridical? Well, O gauge train sets are built at a 1:48 scale and the Eiffel tower is 1,063 feet tall, so my model represents the height of the Eiffel Tower veridically if it is 1,063(1/48) = 22.146 feet tall. Even though lengths in the model represent lengths in the world, veridicality does not depend upon any particular length in the model and any particular represented
length being the same. Instead, it depends upon the systematic relationship between relative lengths in the model and relative lengths in the world.

Our perceptual experience, including color vision, is no different. Experience represents the world by measuring the relative values of different stimuli with differences of corresponding relative value on an internal, physiologically determined scale. But this feature of perceptual experience does not impede veridicality, it enhances it. Consider, for example, our perception of lightness. Whether we experience a stimuli as “white”, or maximally light, depends heavily on context. As discussed in Chapter 2, in a pitch black room, only a handful of photons are needed to register an experience of whiteness. On a sunny day, the same number of photons will register as pitch black (or rather, they won’t affect experience of lightness at all). But this isn’t a breakdown in the veridicality of our experience of lightness! It’s a demonstration of the remarkable flexibility of the visual system in registering relative lightness.

But isn’t color somehow special in its relationship to the world, different from, say, length? I think it is a mistake to think here that color poses some special challenge, not to be found in the perception of length and breadth. This is the classic primary / secondary quality distinction. From Democritus on, various philosophers have worried that, if all that exists are “atoms and void” then there is no room for “color” in the physical structure of the world. Although no scientist since Newton has thought that all that exists are atoms and void, regular claims by color scientists seem to imply that color is not a property of objects in the world. There are two important points to make here, which I will discuss in turn. First, the qualitative worries which undermine the straightforward interpretation of color perception as perception of some external property apply equally to spatial perception. Second, claims made by color scientists that color is not a physical property should not be interpreted as claims about the ontology of the world.

One might think that spatial perception is qualitatively different from color perception because spatial experience represents lengths as lengths whereas color experience represents some properties which are not themselves colors (e.g. surface reflectance properties) as colors. This worry motivates the dispositionalist response to the question of color realism: no, colors are not properties of objects, but there
are other properties of objects which cause dispositions in us to perceive particular colors. But this comparison rests on a misunderstanding of our experience of spatial properties. If lengths as we experience them and lengths in the world do not share the same properties, then the close match between experience of the property length and properties in the world does not obtain. Consequently, it does not appear that there is a principled difference between perception of length and perception of color.

The qualitative difference between lengths as experienced and lengths in the world should be clear to all of us. Most obviously, parallel lines are represented as converging in experience, yet we naturally treat external physical space as locally Euclidean. It appears, then, that the geometric structure of our experience of space and that of the external world are quite different. Blumenfeld (1913) performed a series of experiments on this topic. He asked subjects to arrange distant lights such that they appeared equidistant from a particular vantage point (and in a second condition, to arrange them such that they appeared parallel). He discovered systematic discrepancies between our experience of distance and parallelism and the structure of the world. In the 1940’s Luneburg used Blumenfeld’s data to argue that perceptual space is hyperbolic (e.g. Luneberg, 1947), while in the 1970’s, Indow provided a more subtle analysis (e.g. Indow, 1974). By the 1990’s, it no longer appeared that a single geometrical structure could capture all the features of our visual experience of space. Suppes (1995) argues, for example, that “The most important general feature of visual space is that it is context dependent” (37). He attempts to reconcile empirical counterexamples to Luneberg’s analysis (in Foley, 1972, and Wagner, 1985) with the intuitive constraint that visual space has constant curvature, concluding

What we seem to end up with is a variety of fragments of geometric structures to describe different experiments. A hyperbolic fragment perhaps for alley experiments, a fragment outside the standard ones for the Foley experiments, etc. The goal of having a unified structure of visual space adequate to account for all the important experimental results now seems mistaken. A pluralistic and fragmentary approach seems required. (Suppes, 1995, 43)

Further philosophical discussion of these issues can be found in Suppes (2002), Chapter 6, and Hatfield (2003b).
The point of this discussion is just this: the features of our experience of length differ systematically from the features of length in the world. In order to better understand how we experience length, we need to keep these two spaces distinct, i.e. lengths in the world and lengths in experience are not the same thing. The general apparatus for analyzing the relationship between our experience of space and external space is to model each with some geometrical structure (e.g. Euclidean for the external space and hyperbolic for the experience of space) and examine a) experimental evidence for the precise structure of the internal space, b) the exact nature of the mapping between external space and internal experience, and (eventually) c) the physiological basis for this mapping. Yet this is the exact approach we saw in Chapters 2 and 3 for color! Model both the internal and external domains as geometrical spaces, then look at the mapping between them and its physiological basis.

From this perspective, the main qualitative difference between perception of color and perception of space is that, in the case of color, there is a gross reduction in dimension from represented space (infinite dimensional) to representing space (three dimensional). A dimensional reduction has also been proposed for our experience of space, however!

In the eighteenth and nineteenth centuries the majority of theories of visual perception were built upon the view that during the process of vision there occur two conscious states with quite different phenomenal properties. The first state is a mental representation of the two-dimensional retinal image. The second is our experience of the “visual world” of objects distributed in depth. According to the then commonly accepted theory, the mental correlate of the retinal image is the truly immediate component of perception, and it provides the raw material from which the mind generates the three-dimensional visual world. Yet this retinal correlate—the “sensory core” of the perceptual process—typically goes unnoticed, and the percipient takes his experience of the three-dimensional visual world to be direct and unmediated. Although it may seem odd that an unnoticed state of consciousness should be viewed as the psychologically fundamental component of the visual process, that which we have labeled the “sensory core” has played a prominent role in visual theory since Berkeley drew his celebrated distinction between the immediate and mediate objects of vision. (Hatfield and Epstein, 1979, 363–4)
In fact, there is a close analogy here with the analysis of the hierarchy of color experience discussed in Chapter 2. There I argued that the experience of colors in isolation is a lower level in the hierarchy of experience than that of colors in context and corresponds *mutatis mutandis* to a lower level in the chain of neural processing. Likewise, the “sensory core” is a lower level of the perceptual experience of spatial relations which corresponds to a lower level of the physiological processing chain than the experience of space as three dimensional.

**But what about inverted spectra?** Philosophers might worry that color perception is nevertheless still qualitatively different from spatial perception because of the possibility of inverted spectra. Isn’t it possible that the quality of your experience when viewing a blue stimulus and mine when viewing a yellow stimulus are identical and vice versa? In this case, one of us would have an “inverted” spectrum of color experience, in which the qualitative features of my color experiences do not match onto the world in the same way that yours do. Nevertheless, I use color terms to apply to the same stimuli as you do, so we can never discover that one of our spectra is inverted. But this is the essence of the measurement perspective: the “what it’s like” of a particular perceptual experience is *irrelevant* to its representational features. All that is required for color experiences to represent is that there be distinct such experiences, and that these be arranged into a geometrical structure. Change in the quality of these experiences amounts to nothing more than a change of units.

In the case of the perception of space, I see no reason in principle why we can’t apply a qualitatively similar thought experiment. Suppose, for example, my subjective experience of any length is as twice the size of your subjective experience. Nevertheless, we both make precisely the same assessments of relative length. An inch will

---

1Of course, the intuitive plausibility of the spectrum inversion problem depends substantially on the supposed symmetry of the color solid. As we saw in Chapter 4, however, the color solid is actually highly asymmetrical. In order to preserve the discriminative and qualitative features of the color solid while “inverting” hue, one would also have to “invert” the experience of lightness and darkness (c.f. Griffin, 2001), or allow that the inverted spectrum observer can actually experience colors (in particular light shades of purple and dark shades of yellow) impossible for the standard viewer, and vice versa. These sorts of considerations have led Hardin to argue that spectral inversion is actually not possible, or at least that a person who experiences an inverted perceptual experience of colors could easily be detectable due to qualitative differences in his discriminatory behavior (Hardin, 1997).
look twice as long to me as it does to you, but we both agree that an inch is 2.54 centimeters and that a ruler is twelve inches long. The absolute lengths we perceive objects as having are totally irrelevant to representational success; what matters are the relative lengths. So long as relative lengths are preserved, absolute length constitutes a mere choice of units, as between inches and centimeters or Celsius and Fahrenheit. The same holds of the inverted spectrum.

Philosophers frequently remark that color scientists themselves don’t seem to believe colors are properties of objects in the world. Byrne and Hilbert (2003), for example, amass an array of quotes to this effect (3–4). Cast against the backdrop of the color realism debate, this makes color scientists appear to be eliminativists. Eliminativists take color experience to be a mere artifact of our perceptual system. In their view colors do not exist, i.e. nothing in the physical world is colored. Against the backdrop of the present perspective, this insistence merely looks like good terminological hygiene. Consider this representative quote

The results described here . . . suggest that the nervous system, rather than analyze colours, takes what information there is in the external environment, namely, the reflectance of different surfaces for different wavelengths of light, and transforms that information to construct colours, using its own algorithms to do so. In other words, it constructs something which is a property of the brain, not the world outside. (Zeki, 1983, 783 (emphasis in the original), quoted in Byrne and Hilbert, 2003, 4)

Zeki’s article investigates cells in V1 and V4 which respond to wavelengths of light differentially. He finds some cells which seem to correlate with color experience and others which seem to correlate with wavelength. Although these coincide in the case of colors in isolation, they do not in the case of colors in context. This is what inspires Zeki’s claim. The concluding sentences (immediately following the previous quote) are telling:

At what stage of the visual pathways the information provided by the wavelength-selective and antagonistic wavelength input cells is transformed to construct colour-coded and opponent colour cells, and what algorithms the brain uses to undertake this task represents one of the most formidable and exciting problems in neurobiology. (Zeki, 1983, 783)
The basic attitude here is that *information* about wavelength flows through the visual system, which applies *algorithms* to it in order to construct color experience. The main worry expressed in the previous quote seems just to be that if we treat the visual system as *analyzing* color in some straightforward way, we will be mislead because color experience and wavelength come apart in the experience of colors in context. But this is only a challenge to the representational power of color experience if colors represent local absolute wavelength value.

It is telling that the motivation for Zeki’s “eliminativism” has nothing to do with the worry that extension is the only property in the world independent of humans. And rightly so given that the science of today populates the world with forces as well as objects. Rather, it derives solely from the disconnect between wavelength stimuli and color experience. But this doesn’t tell us anything definite about the representational properties of color experience, only that *if* colors represent wavelengths, *then* they appear to do so remarkably inefficiently.

We’ve discussed dispositionalism and eliminativism, what about “realism”? The definitive statement here, perhaps, is Byrne and Hilbert (2003). None of the discussion here *refutes* the arguments given by Byrne and Hilbert. They identify a number of bullets one must bite if one wishes to be a *reflectance physicalist*, i.e. if one wishes to claim that colors are physical properties of objects, and the particular properties identified with color are surface reflectance profiles. Byrne and Hilbert then proceed to bite all these bullets in their defense of reflectance physicalism.

Perhaps the hardest bullet to bite is the epistemological relationship we have with color on the Byrne and Hilbert view. They consider a possible objection from Hardin based on the variations in color assessments across individuals (and for an individual across various viewing conditions). Hardin concludes from this variation that if colors are identified with surface properties, then there must be widespread misperception of color—but this seems to contradict the initial intuitive motivation for color realism, namely that color experience is often veridical. Byrne and Hilbert respond:

We think this argument fails at the first stage, because the conclusion is not unacceptable. . . . note that the conclusion is not especially astonishing or at odds with apparently obvious facts. The conclusion is not that people
rarely see objects as having the colors they actually have, but that they rarely see objects as having the *determinate colors* they actually have. It is consistent with the conclusion that people typically see green objects as green, orange objects as orange, and so forth. (Byrne and Hilbert, 2003, 17)

However, this response combines with another feature of the Byrne and Hilbert view to produce a bizarre consequence. This other feature of their view is that the categories of reflectance types which constitute a color do not have any physically distinguishing feature other than that they are perceived as the same color by humans:

> So the reflectance-types that we identify with the colors will be quite uninteresting from the point of view of physics or any other branch of science unconcerned with the reactions of human perceivers. This fact does not, however, imply that these categories are unreal or somehow subjective. It is just a plain matter of fact that an object has a particular type of reflectance, and this fact need not depend in any interesting way on the existence of creatures with color vision. (11)

So, there is no human-independent test based on physical categories for determining what color a surface is. But we’ve also acknowledged that there is widespread misperception of surfaces, so we can’t use human perception of surfaces to determine unique colors either. Byrne and Hilbert emphasize that this epistemologically unsatisfying state of affairs does not defeat their defense of reflectance physicalism.

> From the fact that we have no good reason to believe, of any chip, that it is unique green, it does not follow that we have no good reason to believe that there are any unique green chips. That would be like arguing that we have no good reason to believe that Professor Plum has been murdered, on the ground that there is no particular person who is clearly the culprit. (17)

(For a challenge to the Professor Plum analogy, see Cohen, 2003.) Saving the most unpalatable bullet-biting for the footnote to this passage:

> [footnote 50] Thus we are prepared to countenance “unknowable color facts”—that a certain chip is unique green, for instance. And so should any color realist who accepts some assumptions that are (we think) highly plausible. (Byrne and Hilbert, 2003, 21)
Now, what have Byrne and Hilbert achieved with this argument? They have saved the claim that physical objects are colored, at the expense of acknowledging that we can never know these colors. If it is important to one to insist that physical objects are colored, then Byrne and Hilbert can help them do so in a consistent way. However, the purpose of this section has been to emphasize repeatedly that only by keeping the represented properties of the physical world and the representing properties of experience distinct can we investigate the relationship between experience and the world in a nuanced and detailed way. The consensus view (amongst color scientists) appears to be that color terminology should be applied to representing experiences rather than the world. Do we gain anything, either in terms of conceptual clarity or scientific insight, by arguing that color terms can consistently be applied to surface reflectance properties? I don’t think so.

In fact, this appears to be the view of color scientists as well. In a response entitled “Color: A Vision Scientist’s Perspective”, Davida Teller makes the same argument I have presented here:

Now, as far as I can see, color realism is the view that of the vision scientist’s three entities—surface spectral reflectance, neural signals, and perceived color—one is color, and the other two are not. But if you ask a color scientist which of the three entities is color, she will answer that the question is ill-posed. We need all three concepts, and we need a conceptual framework and a terminology that makes it easy to separate the three, so that we can talk about the mappings among them. Color physicalists can call surface spectral reflectance physical color if they want to, although surface spectral reflectance is a more precise term. But to call it color (unmodified) is just confusing and counterproductive, because for us the physical properties of stimuli stand as only one of three coequal entities. (Teller, 2003, 48)

Teller also endorses the analysis offered earlier of how to read the many quotes from color scientists purportedly showing they are eliminativists about color:

...in a sense, we line up our usage with the subjectivists rather than the physicalists. But I think the authors of the textbooks quoted by B&H are trying primarily to insist on the distinction between physics and perception, and only secondarily to reserve color terms for perception.
We care much more about our fundamental distinctions than we do about who owns the word *color*. (Teller, 2003, 49)

Finally, from a scientific perspective, the nail in the coffin of the Byrne and Hilbert view is the epistemological issue. If we cannot ever know the unique color of a surface, what is the *point* in saying it has one? Hardin himself draws an analogy here with the electromagnetic ether:

[Discussing footnote 50] There is at least a whiff of ether here, the electromagnetic ether whose undulations were supposed to be the mechanical basis of electromagnetic phenomena. The null result of the Michelson-Morley experiment left one with two choices: Regard the earth’s motion through the ether as an unknowable fact, or else dispense with the ether altogether. Empirical science opted for the second course. Have B&H opted for a chromatic ether? (Hardin, 2003, 32–3)

Again, my point here is not that the arguments of Byrne and Hilbert fail; indeed, they succeed admirably within the context of the problem they have set themselves. My point is that this problem is irrelevant to deepening our understanding of the world, both from the perception as measurement perspective, and from the perspective of color science. I happen to believe (and hope to have shown in the previous 5 chapters) that the perception as measurement perspective fits especially closely with the practice of scientists. However, even if one rejects the overall framework I have defended in this work, if one is to be serious about understanding perception, one must acknowledge that

1. My experience of the world and the world itself are distinct.

2. My experience of the world (even in the case of color) tracks the world in some systematic way.

Both these facts are acknowledged by everyone who thinks seriously about perception (including Byrne and Hilbert and all participants in the color realism debate). But surely fact 1 indicates that it would be a mistake to conflate properties in our experience with properties in the world, less we run the risk of confusing two distinct
domains. It is in this sense that objects in the world are “obviously” not colored (because “color” as the word is usually used is a collection of properties of the world as we experience it (c.f. Chapter 4)). Yet surely fact 2 implies that there is a systematic informational relationship between experience and the world, and it is in this sense that objects “obviously” are colored (just as noises are loud, electric shocks are painful, and bowling balls are heavy).

Therefore, I think we should ignore the “problem” of color realism. However, this is not meant to deny that there are interesting philosophical questions about color. In fact, I believe rejecting the color realism debate and taking the perception as measurement perspective opens up the door to a whole new world of diverse and subtle philosophical questions about color. For example, given the facts listed above (and the discussion in Chapter 3), some obvious questions to ask are

1. Just what categories in nature does color experience successfully track?
2. How efficient is color experience in tracking these categories?
3. Do these categories constitute interesting natural groupings of relevance to the organism’s survival? Or are they merely artifacts of arbitrary features of the physiology of the organism’s visual system?

The rest of this chapter will suggest some considerations for developing answers to these questions.

6.3 Color Categories: Arbitrary or “Physically Interesting”?

Byrne and Hilbert bite an awfully distasteful bullet when they acknowledge that “the reflectance-types that we identify with the colors will be quite uninteresting from the point of view of physics or any other branch of science unconcerned with the reactions of human perceivers” (Byrne and Hilbert, 2003, 11). This is a natural response, however, to the problem of metamers, or distinct stimuli which produce an identical color
CHAPTER 6. REALISM, METAMERISM, AND BEYOND

experience. Some naturalists are willing to bite this bullet; others, however, continue to search for a “physically interesting” way to characterize metameric groupings.

Paul Churchland, for example, attempts such a characterization because he takes it to be required of a realist response in the color realism debate. Now, on the one hand, I think Byrne and Hilbert have shown that one can consistently be a realist without worrying about whether or not metameric groups are in any sense natural physical categories. On the other hand, my naturalism is not as stringent as Churchland’s; in particular, my naturalism demands only that any theory I develop ultimately derive from descriptive facts about the world. I see no reason to limit myself to physical facts, in the sense of facts in physics or facts about microphysical structure, in particular. So, my naturalism does not require reductions to “the physical”.

However, although I do not endorse either of these motivations for Churchland’s project (nor do I endorse his solution), I think the search for categories in nature which are tracked or identified by our color categories is a worthwhile one—certainly, we should not give up as quickly as Byrne and Hilbert (despite the pessimism of color scientists on this point). For one thing, the existence of such natural categories would go a long way toward explaining the idiosyncratic features of our color experience. For another, our very success in navigating the natural world with our particular color system seems itself evidence that we have latched onto some feature(s) of interest. Figuring out what those features are will help us to better understand ourselves.

As we saw in Chapter 5, natural signs contain information about many different aspects of the causal network in which they are embedded. This general point holds for natural measurement as well. So, another way of phrasing the question is this: We’ve seen in Chapter 3 that color experience measures the world, but what does it measure? Surface reflectance properties are one candidate, but are there more interesting ones? After introducing some of the features of metamers, I will briefly discuss two types of strategy toward solving the problem of metamers, which I call structural and ecological. A structural solution focuses on the space of surface reflectance profiles, and attempts to find some structural feature of that space which will partition it into metameric categories independent of the human visual system. An ecological solution
appeals to features of the environment of interest to humans which themselves cause surface reflectance profiles.

6.3.1 Metamers in Principle and in Nature

If spectral power distributions of light $A$ and $B$ are \textit{perceptually indistinguishable}, then we say they are \textit{metamers} or that they form a \textit{metameric pair}, which we notate $A \sim B$. In the color matching experiment described in Section 2.5.2, any two test lights for which the subject sets the primary lights at the exact same values are metamers.

The notion of a metamer extends from the perception of colors in isolation to the perception of colors in context in a natural way. In particular, we can add context to the setup of the color matching experiment by presenting the bipartite field within a colored surround. In the case of colored lights, metameric matches remain indistinguishable no matter what the context. Use $\beta(A)$ to indicate the subjective appearance of the light $A$ presented in context $\beta$. Then $A \sim B$ implies for any context $\beta$ that $\beta(A) \sim \beta(B)$.

In general, metameric matches do not hold across contexts, i.e. in asymmetric matching experiments. For $A \sim B$ and $\beta \neq \beta'$, $\beta(A) \sim \beta(B)$ and $\beta'(A) \sim \beta'(B)$, but $\beta(A) \not\sim \beta'(B)$. We saw in Section 2.6.2, that the rich contextual environments used as stimuli for asymmetrical matching experiments on color constancy do not necessarily allow for precise matches. The relatively simple contextual matching experiment described here, however, \textit{does} allow for precise matches across contexts. Nevertheless, these matches have different features from metameric matches within contexts. For example, if we increase the brightness of a test light and its metameric match by the same value, they remain perceptually indistinguishable. Using $t \cdot A$ to indicate light $A$ is increased in brightness by $t$ (i.e. multiplication of the vector $A$ by scalar $t$), then $\beta(A) \sim \beta(B)$ implies that $\beta(t \cdot A) \sim \beta(t \cdot B)$. However, this property does not hold for across context matches, i.e. $\beta(A) \sim \beta'(B)$ does \textit{not} imply that $\beta(t \cdot A) \sim \beta'(t \cdot B)$ (see discussion in Suppes et al., 1989, Section 15.4.4).

We can define \textit{surface metamers} by analogy with the light metamers found in color...
matching experiments. However, we must be careful here, for surface metamerism does not have the same properties as light metamerism. In particular, the appearance of a surface does not depend only upon its reflectance profile, but also upon the spectral power distribution of the illuminant. In general, it does not make sense to say of two surface reflectance profiles \( A \) and \( B \) that they form a metameric match. Instead, we must fix the illuminant with respect to which they constitute a metameric pair. Use \( \Gamma(A) \) to indicate the appearance of surface reflectance profile \( A \) under illuminant \( \Gamma \). Suppose \( \Gamma(A) \sim \Gamma(B) \) for \( A \neq B \), then if \( \Gamma \neq \Gamma' \), in most cases \( \Gamma'(A) \not\sim \Gamma'(B) \). And, perhaps more importantly, for any \( A \neq B \), there exists some \( \Gamma \) such that \( \Gamma(A) \not\sim \Gamma(B) \).

Surface metamerism is very important in industrial applications where we would like to control or reproduce color. Suppose, for example, I wish to reproduce a painting as reference in an art book. In general, the pigments used in the printing process will differ radically from those in the original painting. However, I would like to ensure the reproduction in the book matches the appearance of the painting as closely as possible despite the gross physical difference between the causes of color in each case. Since I cannot in general guarantee an exact match across different illuminants (i.e. conditions in which the book might be read), I would like some measure of the degree to which the reproduction matches the original. A complex subfield in the colorimetry literature deals with effective strategies for measuring degree of metamerism in just such applications (see e.g. Wyszecki and Stiles, 1982, Section 3.3.10).

So, light metamerism remains metameric matches across changes in context, because in the case of the direct perception of light stimuli, the only contextual effect comes from the surround. However, surface reflectance profiles are subject to at least two types of contextual affect. The first is due to the reflectance profile of the surround, the second is due to the spectral power distribution of the illuminant. Since surface metamerism is not invariant across changes in illuminant, surface metamerism is not in general invariant across contexts.

This difference may motivate a different philosophical treatment of surface metamerism and light metamerism. We are worried that metameric categories do not form natural groupings. But metameric categories of surfaces are only measured relative to a light source. If the reflective properties of surface \( A \) and surface \( B \) are different, then in
general we can discover this difference by viewing them in a variety of contexts. This
counterfactual feature of surface metamers may help suggest a solution to the problem
of metamers, but this possibility has been underexplored in the philosophical litera-
ture. One reason is just the dominance of the color realism debate. If one defends
eliminativism about color, then the problem of metamers simply doesn’t arise. If one
defends physicalism or dispositioinalism, then the problem of metamers does arise,
but the distinctive features of surface metamers as opposed to light metamers don’t
help you. The reason is just that the ability to distinguish $A$ and $B$ (metamers with
respect to illuminant $\Gamma$) once they are placed under illuminant $\Gamma'$ depends upon a
difference in the subjective appearance of the two surfaces. But this difference itself
poses a problem for an absolute correspondence between color experience and surface
properties. On the relative measurement view defended here, this phenomenon is not
so obviously problematic.

A great deal of effort has been put into generating metamers artificially for in-
dustrial applications (for printing, paints, dies, monitors, film, etc.). Most examples
of metamers found in textbooks have been calculated from our theoretical model of
a standard observer based on color matching experiments. For example, Figure 6.1
depicts twelve metameric reflectance profiles, each of which should produce an ex-
perience of “hypothetical grey” in the CIE standard observer if viewed under CIE
standard illuminant C. These profiles were generated by first finding solutions to
three simultaneous equations ensuring the same tristimulus values are generated,
then scaling the results to ensure they are all “realistic” (i.e. do not assign more than
100% reflectance to any particular wavelength (Wyszecki and Stiles, 1982, 188–91).

Even a casual glance at Figure 6.1 should serve to illustrate the essential problem.
The twelve curves presented bear no obvious similarity to each other whatsoever.
They have peaks scattered around the spectrum. Some curves have only low broad
humps, while others have sharply spiked peaks. Some curves almost look like “oppo-
sites” when graphed in this fashion. Consider curve 1 and curve 12 on the bottom
graph. Around 480 nm curve 12 has a sharp peak, while curve 1 has a low trench;
curve 1 peaks strongly around 550 nm and somewhat more weakly around 420 nm,
correspondingly curve 12 dips in both these regions; then, around 700 nm, curve 12
Figure 6.1: Twelve metameric surface reflectance profiles for the CIE 1931 standard observer and CIE standard illuminant C. (Wyszecki and Stiles, 1982, Figure 7(3.8.2), reproduced by permission of John Wiley & Sons, Ltd.)
exhibits a gentle rise, while curve 1 sits at zero. How can all these curves constitute a natural category?

An interesting empirical desideratum here is that metameri c pairs may be quite rare in nature. (Of course, to actually measure whether or not this is the case would be a mammoth task, but some of the considerations which support this conclusion will be discussed below.) Suppose, as many color researchers suspect, metameri c pairs are quite rare in nature—is this a blessing or a curse for the naturalist? On the one hand, if metameri c pairs are rare, then the possibility of metameri c matches shouldn’t pose a barrier to the visual system’s veridical representation of those surface properties which do appear frequently in a natural environment. On the other hand, if our strategy for cashing out color content is in terms of metameri c categories, then how did color vision evolve to latch onto such categories? What would be the basis of the selective pressures for distinguishing them? (This argument is offered against Churchland’s solution to the problem of metameri in Kuehni and Hardin, 2010, 91.)

In the following two sections I examine some suggestive strategies for providing human independent, naturally motivated characterizations of metameri c categories. The discussion will focus on surface metameri s because this topic has generated much more philosophical and scientific interest. Also, I believe, there are promising strategies which present themselves in the surface case but do not appear well motivated for the case of light metameri s.

6.3.2 Structural Solutions

One strategy for solving the problem of metameri s is to look for some structural feature of the space of possible surface reflectance profiles with respect to which metameri c categories appear natural. In both scientific and philosophical circles, there is widespread pessimism that any such structural features exist. This is the motivation behind Byrne and Hilbert’s ready rejection of the idea that the categories of reflectance properties they identify with colors are in any way “physically interesti ng”. Nevertheless, there have been attempts to discover such structural features within the space of possible reflectance profiles, two of which we will examine here.
Contra Byrne and Hilbert, Paul Churchland (2007) has recently argued that metameric categories are “physically interesting” by defining a transformation of the space of possible reflectance profiles which does not depend upon features of human vision. He then argues that the perceptual experience of color represents this new space, which he calls the space of “CA-ellipses”, and that there is a straightforward homomorphism from this space into that of possible color percepts. So, the strategy involves two mappings, each of which must satisfy a crucial desideratum:

1. A map from surface reflectance profile space into the space of CA-ellipses which is “physically interesting”

2. A map from CA-ellipse space into color percept space which preserves color categories

As it turns out, both of these steps fail. Nevertheless, it may be helpful to briefly examine Churchland’s strategy as it is intuitively suggestive.

The space of possible surface reflectance profiles is infinite dimensional. It assigns some percentage of reflectance (i.e. a value in [0, 1]) to each wavelength within the visible portion of the electromagnetic spectrum, roughly the interval from 400 nm to 700 nm. We can represent profiles within this space as functions $f : \lambda \rightarrow [0, 1]$, where $\lambda$ can take on any value in the range 400–700. Intuitively, such functions are curves on a finite two-dimensional surface, where one axis represents efficiency of reflectance and the other represents wavelength (Figure 6.2.a).

Churchland proposes transforming the space in Figure 6.2.a by wrapping the plane into a cylinder (Figure 6.2.b). Next, he points out that any curve on the surface of this cylinder has a unique approximation, which he calls the “canonical approximation ellipse” (CA-ellipse, e.g. Figure 6.2.d), defined as the curve formed by intersecting the cylinder with the unique plane satisfying two constraints. Call the area of the region above the plane but below the original curve $U$ and the area of the region below the plane but above the original curve $D$. Then the approximating ellipse is “canonical” if it satisfies:

1. $U = D$
2. The tilt of the plane from the horizontal is such that it minimizes $U + D$

Next Churchland considers the space of possible CA-ellipses. This space can be defined by three axes: rotational position is a circular axis defining the orientation of the intersecting plane’s tilt; tilt angle can be treated as a radial axis; and the altitude of the intersecting plane in the center of the cylinder can be captured with a vertical axis (Figure 6.2.e). Churchland is inspired by the similarity in structure between the space of possible CA-ellipses (Figure 6.2.f) and the spindle-shaped space of possible color percepts (e.g. as in Figure 2.9) and claims that a straightforward homomorphism exists from the first into the second.

Suppose Churchland’s analysis works, would it constitute a solution to the problem of metamers? Churchland himself thinks so. He describes the similarity in structure between the space of CA-ellipses and the space of color percepts as “a color realist’s prayer” (133). More specifically, the map from CA-ellipse space into the color solid differs from a direct map from surface reflectance profiles into the color solid in that the first mapping preserves betweenness relations and the second does not. Furthermore, he takes his approach to differ from the Byrne and Hilbert view precisely in that he “can indeed specify, in terms ‘of interest to a physicist,’ the feature that unites the family of metamers for a given commonsense color: they all share the identical reflectance-space ellipse as their canonical approximation” (Churchland, 2007, 147)

As it turns out, both steps in Churchland’s argument have been convincingly refuted. Wright (2009a) argues that Churchland is incorrect to claim that CA-ellipses are “of interest to a physicist”:

CA ellipses are not physical, at least not in a way that is helpful to Churchland’s desired brand of physicalism. While CA ellipses are generated by what can be granted are ‘objective means’—all that happens is a mathematical transformation on a representation of something that is clearly physical—that is no guarantee that the resulting construction has any interpretation that is natural from the standpoint of physics. (Wright, 2009b, 395)

Wright analyzes meaningfulness in physics in terms of invariance under physically permissible transformations. He points out that from the standpoint of physics, the
Figure 6.2: a) a reflectance profile; b) the same profile rolled into a cylinder; c) the sine curve corresponding to profile (a)’s CA-ellipse; d) the CA-ellipse formed by a planar intersection; e) the three dimensions defining a CA-ellipse; f) the space of possible CA-ellipses. (Churchland, 2007, Figures 4 (127), 5 (129), and 6 (131), reproduced by permission of the Philosophy of Science Association)
choice of 400 nm and 700 nm as the ends of the region to consider is arbitrary.\textsuperscript{2} However, changing the end points of the region which one wraps into a cylinder can quite dramatically affect the orientation of a CA-ellipse. Consequently, CA-ellipses are not invariant under a physically permissible transformation, and thus are not physically meaningful (Wright, 2009b, 396–9).

Kuehni and Hardin (2010) emphasize the failure of step 2 in Churchland’s argument, the mapping from the space of CA-ellipses into the perceptual color solid. The arguments presented are completely convincing. After emphasizing that Churchland fails to distinguish a unique contemporary model of perceptual color space (e.g. the CIELAB space discussed in Chapter 4, Section 4), they test the metameric categories predicted by Churchland’s CA-ellipses against sets of metameric curves generated by the same method which produced Figure 6.1. They discover that Churchland’s claim is simply wrong, and metameric pairs do not in general receive the same CA-ellipse. Since Churchland’s method cannot predict color matches (anywhere near) as accurately as procedures which make use of facts about cone cell response, he has failed to provide a serious human-independent candidate for color categories.

The machinery [for calculating metomers] is comparatively simple and the prediction of matches (equal color appearance) requires only a simple mathematical model. The corollary of linear relationship between the results based on different primaries, say cone sensitivity functions and different versions of color-matching functions, has been extensively empirically substantiated. A claim of predicting human visual metomers without the aid of cone functions requires an equal level of reliability. That reliability has not been demonstrated by Churchland. (Kuehni and Hardin, 2010, 91)

In my view, the arguments of Kuehni and Hardin are significantly more damning than those of Wright. Had Churchland’s analysis worked, it would indeed have

\textsuperscript{2}Perhaps it is even arbitrary from the standpoint of the visual system itself. In Chapter 2, Figure 4 we can see that cone cells do respond to wavelengths lower than 400 nm and higher than 700 nm, although our strongest perceptual responses are produced by stimuli falling squarely within the 400 – 700 nm range.
constituted at least a step forward on the problem of metamers, even if his definition of CA-ellipse space failed to be invariant under meaningful physical transformations. Had the mapping from CA-ellipse space into perceptual color space succeeded, then CA-ellipse space could have provided evidence for patterns to look for within metameric categories which themselves may eventually have led to a more physically interesting characterization of these categories.

In fact, there are other approaches to finding patterns within metameric categories. One small and puzzling literature concerns the number of crossings of metameric curves. Stiles and Wyszecki (1968) demonstrate that if two spectral reflectance curves are metameric, then they must intersect at least once. Furthermore, there are constraints on the number and location of the intersections between curves in a metameric category. This work is not enough to determine metameric matches, but it does indicate patterns in the relationship between metameric profiles.

Of course, the constraints defined in Stiles and Wyszecki (1968) are not in any sense “physically natural”, but, just as with Churchland’s project, they may provide evidence for more natural representations of surface reflectance space. In fact, Thornton (1973) made the much stronger claim that metamer intersections always occur at 448 ± 4 nm, 537 ± 3 nm, and 612 ± 8 nm. This would have been a quite interesting constraint on metameric categories were it correct, but Ohta and Wyszecki (1977) argue that Thornton’s analysis depends critically on the mathematical procedure he used for generating metamers. Nevertheless, there are still patterns here.

Figure 6.3 graphs the frequency of intersection at each wavelength for two sets of 990 metamers each. The sets were produced using two distinct methods of metamer generation (graphed separately by the black line and the dashed line). The peaks indicate at least a statistical pattern in the location of intersections. Further patterns were found in the number of intersections per pair, and even mutual constraints between number and location of intersections were found (Ohta and Wyszecki, 1977, see also Wyszecki and Stiles, 1982, Section 3.8.3). Of course, this is a long way from helping us understand what kind of natural groupings might characterize metamers, but an initial step in this direction was taken by Brill (1987), who attempted to empirically check Thornton’s analysis of the location of intersections using statistical
methods and naturally occurring metamer.

6.3.3 Ecological Solutions

This brings us to the second strategy for solving the problem of metamer, which focuses on the conditions which obtain in the world, rather than in principle mathematical models. For example, even if interesting structural constraints on metamer surface reflectance profiles cannot be found, we can search for natural correlates for metameric categories in the chemical structure of objects. One of the most plausible functions of color vision is the quick and efficient location of foodstuffs; yet edibility correlates strongly with the presence of certain chemical compounds, e.g. chlorophyll, rather than the specifics of a particular reflectance profile. Perhaps the correlates of color categories are not to be found within the space of spectral power distributions, but within the space of chemical causes of spectral power distributions.

This is the ecological perspective, which searches for features of the environment of interest to the organism and posits that the organism represents these features. The figure most strongly associated with the ecological movement is J. J. Gibson (1966, 1979). By focusing on surface reflectance profiles and spectral power distributions, we have used our knowledge of the nature of the physical interaction at the retina.
as a clue to that feature in the world which the experience of color represents. The ecological perspective suggests we look to more distal features of the causal chain from world to experience.

If we take the ecological perspective, one apparently arbitrary feature of the visual system can immediately be explained. Why do we only perceive light in the range 400 – 700 nm? Isn’t this just an arbitrary segment of the enormous electromagnetic spectrum? As it turns out, there is a quite convincing response to this question, due to Kurt Nassau:

> It might seem remarkable that so many distinct causes of color should apply to that small band of electromagnetic radiation to which the eye is sensitive, a band less than one “octave” wide in a spectrum of more than 80 octaves. So much happens in this narrow band because this is the region where the interaction of radiation with electrons first becomes important. Radiation at lower energies induces relatively small motions of atoms and molecules, which we sense as heat, if at all. Radiation at higher energies has a destructive effect since it can ionize atoms, that is, completely remove one or more electrons, and can damage molecules permanently. Only in the narrow optical region to which the human eye is sensitive is the energy of light well attuned to the electronic structure of matter, with its wide diversity of colorful interactions. (Nassau, 2001, 31)

On the one hand, this consideration is extremely powerful as an explanation for why the visible range of the spectrum is 400 – 700 nm. On the other hand, these considerations also lend credence to the idea that the function of color experience is to determine the chemical structure of distal objects. If we were only interested in the location of objects, say, then we would not need to restrict our attention to just those wavelengths which can interact with surfaces enough to respond differentially to their chemical structure, but not enough to actually damage it.

The traditional ecological approach employs task analysis to generate hypotheses about evolutionarily selected function. As such, the traditional approach fits more closely with the teleosemantic approach than the information theoretic one I defend here. One example of an ecological analysis of color content is Hatfield (1992). Hatfield argues that one function of color vision is the discrimination of interesting
objects in the organism’s environment. So long as the system succeeds at this task, the theoretical possibility of metameric matches is essentially irrelevant.

Assume that one function of color vision is to enhance the discriminability of objects and surface features, and that a particular color system serves to promote the discrimination of healthy green plants from soil and rocks. Such a color system must be able to discriminate the surface reflectances of green plants from other reflectances. In evaluating the proficiency of the system, it would be of no consequence if there were physically possible but not actual (nonplant) metameric matches to green plants that the system could not discriminate. As long as such potentially equivocal stimuli were not extant in the environment, the fact that the color system could not discriminate them would not imply a functional deficiency. (Hatfield, 1992, 497)

Hatfield’s goal is to challenge the interpretation of the visual system as a physical instrument for measuring spectral power distributions incident on the retina and with it the idea that metamerism “entails a loss of information”. He concludes:

On this conception, the visual system is not a physical instrument for recording the values of the proximal stimulus as described in physical optics. Rather, it is a perceptual system with the function of representing surfaces as an aid to detecting food and other significant objects. Extended to the case of color vision, this approach suggests that metamerism need not entail a loss of information. If color vision has the function of discriminating particular environmentally given classes of object surfaces, the mere possibility of metamerism may be irrelevant to an assessment of its performance. Further, environmentally extant metamers need not entail a discriminatory deficiency if their discrimination would not yield a biologically significant partition of environmental surfaces. The representational content of color perception might best be conceived in terms of partitions of object surfaces into discrimination classes that are conjoined with adaptively significant objects, and not in terms of a physical specification of spectral energy distributions. (Hatfield, 1992, 501–2)

Hatfield is simply wrong in his claim that “metameric matching need not entail a loss of information”—information loss is an analytic consequence of metamerism, since
it defined as the indistinguishability of physically distinct signals. Since a distinction is erased in metamer matching, information is lost. However, if rephrased, Hatfield’s essential point is correct: loss of information about more proximal stages in the causal chain does not entail loss of information about more distal stages. This point plays a crucial role in Dretske’s definition of primary representation which will be explored in the following section. Here, Hatfield claims that the informational loss found in laboratory conditions and theoretical models of metamerism may be irrelevant in natural environments. In particular, to the extent that we succeed in the ecological task of discriminating surfaces of interest, no information relevant to this task is lost in the causal interaction between experience and world.

Hatfield motivates his discussion with an ecological task analysis and an assumption about the role of selective pressures. In the categorization scheme of Figure 1.1, Hatfield is an “ecological teleosemicist”. However, the hypothesis he advances can be interpreted in straightforward empirical terms, without any need for teleology: is the human visual system adequate to discriminate the actual surfaces found in a natural environment? Once we ask this question, we allow for constraints in the physical structure of the environment to assist in the flow of information from surface to experience. As it turns out, there is some evidence that this claim is correct.

Remember, metamer surface reflectance profiles are generated by mathematical procedures which are not sensitive to any constraints in the physical world. Not all metamer profiles calculated by such procedures are physically realizable. We saw in Figure 6.1 that such profiles can be sharply spiked around narrow frequency bands. In general, however, reflectance profiles found in nature tend to have relatively flat, smooth curves. It appears that the complex physical interaction between light and surface may generate relatively simple constraints restricting natural reflectences to smooth curves. For example, Lythgoe (1979) points out that the reflectance curves of many natural surfaces are straight lines (31), and MacAdam (1981) argues that most biological pigments are essentially just cutoff filters (wavelengths on one side of a certain value are absorbed, on the other side reflected) (128). (These claims are cited in Maloney, 1986, 1674.)

When analyzing the spectral reflectance curves of the Munsell color chips for
other purposes, Cohen (1964) discovered that his data did not satisfy the expected independence assumptions. He performed a linear component analysis and discovered that relatively few components were needed to recover the reflectance curves with an enormously high degree of fit:

Component I extracted 92.72% of the cumulative variance, component II 97.25% of the cumulative variance, component III 99.18%, and component IV 99.68%; considering the reduction in rank, this is one of the highest extractions of variance on record. It follows that three scalar multipliers, $M_1$, $M_2$, and $M_3$, may be assigned to each chip. When the first three components of Table 1 are weighted by their respective multipliers, for any chip, and each row summed, a reconstructed reflectance curve of high accuracy may be obtained for that chip. ...It is to be emphasized that these data are entirely physical; the three scalar multipliers do not correspond to the hue, value, and chroma of psychological specification. Given the psychological specification of a Munsell chip, the chip’s reflectance curve cannot be derived; however, given the three scalar multipliers, the entire reflectance curve can be predicated with high accuracy. (Cohen, 1964, 369)

Cohen started with the reflectance profiles of Munsell color chips. He used statistical techniques to find curves (basis vectors) which, when weighted appropriately and summed, produced approximate matches to the Munsell chip reflectance profiles. What’s surprising here is how few basis vectors are needed to produce a very high degree of fit. Over 99% average fit is achieved with only three basis vectors. Consequently, with over 99% efficiency, a Munsell chip’s reflectance profile can be coded by a set of three scalars, three numbers characterizing the weights of the three basis vectors required to generate the specified reflectance.

This result is fascinating because the human visual system performs precisely this sort of analysis on the incoming color signal, as discussed in Chapter 2. The response patterns of the cone cells act as basis vectors and the degree of response of each cell for a given input is equivalent to the weighting on the corresponding basis vector. As we saw in Chapter 2, color matching experiments demonstrate that experience also can be precisely characterized in terms of linear combinations of three basis vectors. So, the fit with human physiology and behavior is amazingly close, although Cohen’s
analysis makes no physiological or psychological assumptions.

Nevertheless, Cohen (1964) is far from completely convincing. The sample investigated is relatively small compared to the space of all possible natural reflectances. Furthermore, why think that the Munsell chips are representative of natural reflectances? Perhaps some selective factor in pigment choice produced the pattern Cohen stumbled upon. However, Maloney (1986) has addressed some of these worries by extending Cohen’s results to a set of 337 natural reflectance profiles collected by Krinov (1947). Maloney concludes that 6 components are needed to recover reflectance profiles within 0.5% of measured value. However, fewer components are needed for minimal color constancy, i.e. the correct identification of surfaces across different lighting conditions.

The results suggest that constraints on possible surface-reflectance functions and the “filtering” properties of the shapes of the spectral-sensitivity curves of photoreceptors can both contribute to color constancy. (Maloney, 1986, 1681)

What exactly is the source of these strong constraints on natural reflectance profiles? Maloney offers some vague suggestions about the limited types of electron / photon interactions responsible for absorption (1677–8). The statistical character of these interactions for specific photons implies that in general (averaging over many photons), we can treat absorption properties as defining band pass filters. Likewise with the absorption properties of photopigments in the retina (1679–80). Viewed in this way, the inverse problem of recovering a small set of bandpass filters from only the filtered signal seems much more manageable than recovering an infinite dimensional surface reflectance profile. Nevertheless, the discussion here is only suggestive and more work needs to be done to make this treatment fully convincing.

So, from an ecological perspective, the visual system may latch on to natural categories in the world, despite the phenomenon of metamerism. Success in performing this task may depend heavily on natural constraints, the result, perhaps, of simple emergent features of the complex physical interaction between electrons and photons at the surface of an object. Evolutionary analysis in the ecological tradition can
lead us to this conclusion. Statistical analysis of reflectance profiles in natural environments supports it as well. In the next section, I suggest a general strategy for deriving the intentional content of perceptual experience from statistical regularities in the physical environment.

### 6.4 Minimizing Cross-Entropy to Find Intentional Content

A color percept contains information about many stages in the causal chain from distal features of the environment to color experience. It even contains some information about internal features of the neural processing of the color signal. Figure 6.4 provides a very rough schematic of this causal chain. If the informational content of a color percept \( p \) is given by its \( s \)-vector (see Chapter 5), then at least some events at each stage of this causal chain will receive non-zero informational values at the corresponding position in \( s \)-vector\((p)\).

I propose that the *intentional content* of the color percept \( p \) be derived in two steps. First, distinguish that stage in the causal process about which color percepts contain the most efficient information. This stage will constitute the space of variation to which the perceptual experience gives *primary representation* (where representation...
is understood in the sense discussed in Chapter 3). Second, assign as the *intentional content* of *p* that possibility from this particular stage to which *p* assigns the highest probability. The motivation behind step two should be obvious: if I must pick one amongst many outcomes of positive probability to act upon, I pick the one with the highest probability. If we insist on distinguishing some single feature of the world which a color percept is about, then the natural choice is the feature to which it assigns the highest probability.

The motivation behind the first step is just this: distinct events may contain more or less information about different stages in the causal path. In a highly idiosyncratic lighting condition, say an entire scene of unfamiliar objects is illuminated by a single-wavelength red light, we may learn more about the lighting arrangement from our color percepts than about the surfaces. Such idiosyncratic cases may produce an inhomogeneous story about the representational features of color experience. But a story in which some color percepts represent surface properties and other color percepts represent light sources is deeply unsatisfying. Furthermore, it contradicts the perception as measurement perspective. If we view color perception as measurement, then the domain of possible color percepts *as a whole* will represent some domain of mutually exclusive possibilities in the world *as a whole*. Such a domain of mutually exclusive possibilities can be found by singling out one stage in the causal chain from world to experience. The remainder of this section will focus on a solution to this problem.

Dretske (1981) also worried about how to single out one stage in the causal history of a signal as that to which it gives *primary representation*. He recognized that a signal may fail to contain (much) information about proximal stages in its causal history while containing (more) information about more distal stages in that history. In Figure 6.5, for example, the event that *c is B* may lead to the signal *S* by two causal pathways, one leading through the event *d is E*, the other through the event *d is F*. Remember that in Dretske’s theory, *S* only contains the information that *o is P* if *o is P* holds with probability 1 conditional on the signal *S*. In this case, by Dretske’s lights, *S* contains the information that *c is B*, but it does not contain any information about the intermediary causal stage. Within the probabilistic framework developed
in Chapter 5, $S$ contains more information about the event at the source than about the more proximal events.

[Notice that the analysis in the previous section assumes something like this to be the case for metamers. Sure a color percept may have been triggered by any one of a large set of metameric surfaces (the surface $d$ may have reflectance profile $E$ or reflectance profile $F$). But it may be the case that a more distal stage in the processing chain ($c$ is $B$) or a more efficient characterization of metameric groups (i.e. discovery of a $B = E \cup F$) might serve as the representational target of color percepts $S$. This was just Hatfield’s point—loss of information about surface reflectance profiles may not entail loss of information about some more salient feature distinguishing surfaces in a realistic environment.]

But Dretske also realized that $c$ is $B$ itself might follow with probability 1 from some previous event, say that $f$ is $G$ (Dretske, 1981, 159–60). He considers the example of a doorbell. Dretske argues that the relationship between the depression of the doorbell button and the ringing of the doorbell is lawlike and thus satisfies his stipulation that informational relationships hold with probability 1. Yet we don’t “hear” the depression of the doorbell, we “hear” the ringing of the bell. In order to explain this, Dretske introduces the concept of primary representation:

$S$ gives primary representation to property $B$ (relative to property $G$) = $S$’s representations of something’s being $G$ depends on the informational relationship between $B$ and $G$ but not vice versa. (Dretske, 1981, 160)
In Chapter 1, I quoted Dretske’s discussion of this definition without comment. The point is relatively straightforward, however. My auditory experience carries information about the presence of a visitor at the door because of the informational link between visitors and doorbell rings. The information it carries about the ringing of the doorbell itself, however, does not depend upon this informational link. Therefore the auditory experience gives primary representation to the bell, not to the presence of a visitor. Armstrong (1983) finds Dretske’s analysis of this issue one of the most convincing parts of Dretske (1981).

Unfortunately, Dretske’s analysis does not deliver a plausible conclusion if we extend it to the probabilistic framework developed in Chapter 5. Dretske’s stipulation that information only flows when conditional probability equals 1 allows him to rule out proximal stages in a causal chain resulting in experience. The probabilistic framework identified in the previous chapter, however, will assign non-zero informational content concerning these stages to a perceptual experience. For example, consider an experience of color in context. This experience may be triggered by a number of different spectral power distributions of light incident on the retina (this follows from the phenomenon of color constancy, see below). If we apply Dretske’s definition, then we can see that it assigns primary representation to this spectral power distribution even though there is a higher correlation between the color experience and surface properties in the world. This is just because the information my experience delivers about the light incident on the retina does not depend upon the informational relationship between surface properties and light incident on the retina (but the information my experience delivers about surface properties does depend upon this relationship). Perhaps we can recover the spirit of Dretske’s analysis, however.

In order to isolate the single step in the causal chain about which a perceptual experience contains maximal information, we need some asymmetric measure of informational distance. Here, just as Dretske himself, we let safe inferential procedures guide us. Dretske requires that $A$ contain information about $B$ only if $P(B|A) = 1$ because, if the conditional probability is 1, then the inference is sufficiently “safe” for his purposes, namely the preservation of knowledge. Anything less than probability 1 and knowledge cannot flow (see the discussion in Dretske, 1983, especially 84). In
our case, we look again to the “safe” inference procedure suggested by information theory, minimizing cross-entropy.

The remainder of this section will sketch an approach that derives intentional content from the minimization of cross-entropy. This approach uses both probabilities and probability distributions over probability distributions. It is worth emphasizing again, however, that once we set a spatiotemporal region of relevance (say, the environment within which an organism grows up, or the spatiotemporal region within which it evolved), these probabilities (and even the probabilities of probabilities) can all be cashed out strictly in terms of relative frequency. So, our naturalistic desideratum, that our analysis of content be derived strictly from descriptive facts about the world, is satisfied.

Let’s begin with the simpler case of the content of a particular percept. In order to avoid confusion with later definitions, we’ll call this the immediate content. Here is the question: what event or state in its causal past does a particular color percept represent in situ? Let’s consider some potential candidates in Figure 6.4, e.g. the light source, the surface reflectance profile, or the chemical composition of the object.

In order to apply the principle of minimizing cross-entropy, we need to be able to compare probability distributions over a single space. The natural candidate for such a space is the locus of transduction (see Section 2.4.1, i.e. the cone response in the retina. Once the informational signal enters the body, it is subjected to extensive processing ultimately resulting in conscious experience. By hypothesis, the point of this processing is to extract information of interest from the signal at the retina. Also by hypothesis, the source of this information of interest is somewhere in the percept’s causal past. So, the common ground for both representing and represented can be found at precisely that barrier where external information is converted to internal information, i.e. the locus of transduction.

Now, a particular experience of color in context will induce a probability distribution over the space of possible cone activations. This is because, depending on the context, different patterns of activation might result in the same percept. This follows from the phenomenon of color constancy, i.e. my ascription of the same surface color across changing lighting conditions (and thus changing proximal stimuli
Figure 6.6: Schematic of inferential problem for determining the step in the causal chain most efficiently coded by color experience. $T$, $S_1$, $S_2$, and $S_3$ are all probability distributions over the space of possible cone cell activation determined by relative frequencies within a set spatiotemporal region.

at the retina). Likewise, any surface in the environment will induce a probability distribution over cone activation patterns. This relationship is statistical because the surface may be illuminated by different spectral power distributions of light. The relative frequency of different light sources in the environment will then determine the probability distribution over cone activations associated with a particular surface reflectance profile. Similar considerations apply to light sources themselves and possible chemical structures of objects. The situation is depicted schematically in Figure 6.6.

Take $\langle \Omega, \mathcal{A} \rangle$ to be the set of all physically possible cone activations with a suitable algebra defined over them. Then $T \sim p$ is a random variable representing the target probability distribution over possible cone activations induced by a particular color percept. $S_1 \sim q_1$ is a random variable representing the probability distribution over cone activations due to a particular light source; $S_2 \sim q_2$ is a random variable representing the probability distribution over cone activations due to a particular surface reflectance profile; and $S_3 \sim q_3$ is a random variable representing the probability distribution over cone activations due to a particular chemical composition of an object.
What does our percept represent? If it represents what about which it contains the most information, then a reasonable suggestion is that the percept represents that $S_i$ ($i \in \{1, 2, 3\}$) which minimizes the cross-entropy from $T$.

**Definition 6.4.1** Given a space of sensory neuron activation $\langle \Omega, A \rangle$, a random variable $T \sim p$ over that space (the target percept) and a set of $n$ random variables $S_i \sim q_i$ for $0 \leq i \leq n$ (the candidate stimuli), then the immediate content of $T$ is that $S_j$ which minimizes $D(q_j||p)$.

This definition identifies the immediate content of a particular color percept with the particular state in its causal past about which it contains the most information. Remember, however, that we are interested in the relationship between color experience as a whole and some domain of variance in the world as a whole. This is because a representational relationship defined in terms of measurement holds between such domains, not individual events.

We still need to keep our attention on the space of cone activations $\langle \Omega, A \rangle$ in order to make direct comparisons between experience and the world. A single color percept is represented by a probability distribution $p$ over this space. Now consider all possible color percepts; this is a family of probability distributions $p_i$ over $\Omega$, each of which represents a particular percept. But once we consider this family, we should consider the fact that different color percepts are tokened with different frequencies. The relative frequencies of color percepts induces a probability distribution over the family of $p_i$’s. Call this distribution $\mathcal{P}$. $\mathcal{P}$ is the representation of color experience space in the space of cone activations. It is a probability distribution over probability distributions which captures the complete statistical relationship between color experience and the space of cone activation.

Following a similar procedure we can define $\mathcal{L}$, the representation of light source space in the space of possible cone activations, $\mathcal{S}$, the representation of surface reflectance space, $\mathcal{C}$, the representation of the space of chemical structures, etc. Now that we have represented both the space of color experience and the possible spaces which it measures in a single domain, we can again apply the principle of minimizing cross-entropy to determine that domain to which $\mathcal{P}$ gives primary representation:
Definition 6.4.2 Given a space of sensory neuron activation \( \langle \Omega, \mathcal{A} \rangle \), a probability density over distributions on that space \( \mathcal{P} \) (the representing space), and a candidate set of probability densities over distributions on that space, \( \{ \mathcal{L}, \mathcal{S}, \mathcal{C}, \ldots \} \), (the potentially represented spaces, each of which is associated with a different stage in the causal chain resulting in \( \mathcal{P} \)), then \( \mathcal{P} \) gives primary representation to that \( \mathcal{X} \in \{ \mathcal{L}, \mathcal{S}, \mathcal{C}, \ldots \} \) which minimizes \( D(\mathcal{X}||\mathcal{P}) \).

Given this definition, we can define the intentional content of a particular percept in terms of the primary representation of the space.

Definition 6.4.3 Given a space of sensory neuron activation \( \langle \Omega, \mathcal{A} \rangle \) and probability densities \( \mathcal{P} \) and \( \mathcal{Q} \) over distributions on that space, where \( \mathcal{P} \) gives primary representation to \( \mathcal{Q} \), the intentional content of a particular percept \( p \in \mathcal{P} \) is just that \( q \in \mathcal{Q} \) which minimizes \( D(q||p) \).

The success of these definitions is an empirical question. If we examine a plausible spatiotemporal region and count the relative frequencies of lighting, surface, and chemical states within that region, will the application of the above definitions deliver an intuitively plausible analysis of representation and content? I believe it would be a worthwhile project to discover if so, though one which would be, I’m afraid, prohibitively impractical.

Ideally, the difference between immediate content and intentional content will match with intuitive situations in which the two come apart. Consider again the case of a room filled with unfamiliar objects, illuminated by single-wavelength red light. Intuitively, my experience of color in context is still an experience of surfaces as colored. But in this situation I receive hardly any information about surface properties and a great deal of information about the light source. The intentional content of my experience is of the surfaces having a particular reflective property, say, because my experience of colors in context as a whole gives primary representation to surface properties. Nevertheless, in this situation, the immediate content of my experience is just that the light is red since this is the environmental feature which is most plausible given my perceptual experience (i.e. its representation is minimally distant from the representation of my experience).
The beautiful feature of these definitions is that they could be interpreted as providing an analysis of representational function which does not depend upon teleological just-so stories. Rather than argue from a task analysis what the selective pressures should have been on an evolving representational system, we can look at statistical patterns in the environment in which an organism evolved to determine what its representational systems do represent. Working backward, we can look for adaptive advantages to representing these features of the environment in order to understand how selective pressures might have brought this representational system about. But our attribution of representational function does not depend upon these selective pressures, instead we derive evidence about these pressures from our functional analysis.

Of course, another interesting feature of these definitions is that they depend upon relative frequencies in the environment. As these frequencies change, representational success changes as well. For example, our experience of color in context may have evolved in order to represent nutritive features of the environment (recall again the Hanunóo color terms which entangle hue with succulence vs. desiccation), yet they no longer do so in the world of artificial pigments we now inhabit. To me, this is a plausible just-so story which the definitions above effectively turn into a precise empirical question.

6.5 Conclusion

This project developed an account of informational content in nature and applied it to perceptual experience with a special focus on the perception of color. I argued that perceptual experience is hierarchical (Chapter 2) and that any particular level within this hierarchy should be understood as measuring some space of variation in nature (Chapter 3). The semantics of concepts and terms should be understood as deriving from the rich continuous space of perceptual experience by partitioning it into discrete subspaces; this explains how we can token concepts and terms in error, i.e. detached from the direct causal connection between experience and world (Chapter 4).
Contra philosophical lore, we discovered that information theory suggests a natural formalization of informational content (Chapter 5). But informational content is profligate, and any event in nature (including a perceptual experience) will in general contain information about many different stages in the causal chain in which it plays a part. I suggested that we might associate the intentional content of an experience with the highest probability event at that stage in the causal chain which it measures most efficiently. A strategy for assessing this efficiency of measurement is suggested in the previous section.

The purpose of this project has been to see how far information theory can take us in developing a naturalistic account of intentional content. In order to ensure the project is well-defined, I have suggested specific interpretations of the key concepts “naturalistic” and “intentional”. By naturalistic, I have meant only derived from a descriptive account of the world as offered by natural science. The theory of content offered here is naturalistic in this sense. By intentional I have meant only about one thing (uniqueness) and tokenable in error (detachability). I acknowledge that there may be much more to the general concept of intentionality; nevertheless, providing a naturalistic account of these two features of intentional content is already a worthwhile achievement. Chapter 4 presents a naturalistic semantics for terms and concepts which can be tokened in error, while the present chapter suggests a strategy for deriving uniqueness.

In line with this general purpose, the account of informational content provided here has been completely passive. The elephant in the room throughout has been action. Organisms do not passively absorb information about the world, they interact with it in order to acquire nutrients, breed, and avoid predators. An even richer naturalistic story than that provided here can be achieved by moving from the information theoretic to the decision theoretic. In the context of decision theory, we can balance the informational content of percepts against the ends achieved by acting on the basis of those percepts. The feedback loop between experience and action has been the driving source behind the evolutionary development of the rich representational system of human experience. However far we have been able to come looking only at passive features of information, so much farther may we expect to go once we
shift to the decision theoretic framework. This is an area for further research.

Finally, I have suggested in this chapter that the traditional philosophical question about color, the question of “color realism”, is misconstrued and irrelevant. This is not to say there is no place for philosophers in the attempt to understand color experience. Quite the contrary, I argue that there are many more (and more interesting) questions concerning color which philosophers should engage. To list just a handful, serious philosophical analysis can help

1. Determine the dimensionality of the space of possible surface experiences.

2. Determine those categories in nature which color experience tracks and analyze the efficiency with which it tracks them.

3. Determine precisely those features of color experience which are artifactual and those which are representational.

These questions have been discussed in shadowy and inchoate form within the context of the “color realism” debate. Within the framework of perception as measurement, they can be stated more clearly and precisely, without the baggage of unhelpful terminology from the realism debate. This analysis was shown to agree with the attitude of color scientists themselves. By abandoning the realism debate in favor of more nuanced issues such as these, philosophers of color can conform more closely to scientific attitudes, thereby communicating more efficiently with color scientists and contributing their skills to problematic areas within color science. The most obvious candidate here is the problem of color constancy, with which scientists continue to struggle, and to which philosophers of color have started to turn.
Appendix A

Examples of Natural Signs

A.1 Smoke is a Sign of Fire

What is the causal relationship between smoke and fire? In this section we examine this relationship for the specific case of burning wood. The essential point is that one can have smoke without fire and vice versa. The details of the causal relationship will be different for other examples (e.g. chemical fires), but this essential point places a hard constraint on any realistic theory of natural meaning.

To a naïve eye, it appears that wood “catches fire,” and when it burns it emits smoke. By this folk theory, some reaction in or with wood is the cause of fire, and fire the cause of smoke. However, current theories of the combustion of wood indicate that “wood does not burn directly,” but rather

...first undergoes thermal degradation, or pyrolysis—some of the products of which are combustible gases, vapors, or mists. Under appropriate conditions the products may be set afire and, if enough of their heat of combustion is retained by the wood to maintain the pyrolysis, the burning may continue of its own accord until the wood has been consumed except for inorganic products left as ash. Ordinarily wood is set afire by bringing to bear enough heat to start active pyrolysis, and then applying a pilot flame or other source of high temperature to the combustible gaseous products after they have escaped and become mixed with air. In the absence of a pilot flame, much more heat must be supplied before the pyrolysis products will take fire spontaneously. (Browne, 1958, 1)
So, the combustible gases emitted by a heated piece of wood are a precondition for fire. It is these gases, and not the pre-decomposed wood, which catch aflame. Both pyrolysis and combustion are caused by heat, although pyrolysis may occur without combustion in suitable circumstances.

Technically, all the waste products emitted by a heated piece of wood may be called “smoke”; however, some of these gases will be invisible to the naked human eye, while others visible. In order to understand the relationship between smoke and fire which is of interest to us, we must restrict application of the term “smoke” to only those products of thermal degradation visible to the naked eye. If we do this, it is clear that under the right conditions, we can have smoke without fire and fire without smoke. Smoke will occur without fire if the wood is heated sufficiently to emit large particles of waste, yet, due to lack of oxygen (for example) the combustible particles do not ignite. Fire can occur without smoke if the products of the early stages of pyrolysis (simple hydrocarbons), which are invisible to the naked eye, are then separated from the wood and heated sufficiently in an oxygen-rich environment that they burst into flame.

Of course, the notion of “visible smoke” is not particularly important for a scientific theory of combustion, which focuses rather upon the molecular structure of the various products of thermal degradation. Even “fire” does not play a crucial role in the theory of combustion as it is too crude a concept. Contemporary theories distinguish at least two distinct types of ignition occurring in wood: “glowing” and “flaming,” which interact in a complex and, as yet, poorly understood way (Babrauskas, 2001).

In this example, the informational relationship between the readily detectable categories smoke and fire supervenes on an underlying network of causal interaction involving various types of molecules and heat. This underlying causal network is quite complex and not completely understood, yet this in no way undermines the palpable relationship of natural signification between smoke and fire.
A.2 Spots and Measles

Grice uses the example of “spots” and “measles” to motivate his discussion of meaning. Due to Grice, this example has entered the contemporary literature as a paradigmatic example of natural meaning. Nevertheless, a brief examination of the medical literature on measles indicates that one can have spots without measles and vice versa. Furthermore, there is no theory whatsoever of the causal mechanism which produces spots if one is infected with measles! Certainly, given the complexity of the human body, it is natural to expect that the details of this mechanism are exponentially more complex than those of the previous example.

There is actually an ambiguity in Grice’s example. Does Grice mean by “spots” the external spots, or rash, easily visible on a subject experiencing measles? Or, does he mean “Koplik’s spots,” first described in 1896? Koplik’s spots appear inside the mouth in the early stages of measles, vanishing at the onset of the primary symptoms. Although Koplik’s spots are a much more robust indicator of measles than the external rash, they are not easily perceivable to the casual eye. Koplik’s spots are useful to specialists in determining whether a child known to be exposed to measles (say) has actually contracted the virus, but they are not useful in determining that a child previously thought healthy has contracted the disease. This is not only because the spots must be “looked for” inside the mouth, but because once the easily perceivable external symptoms (e.g. the rash) have developed, the Koplik’s spots disappear.

The relationship between Koplik’s spots and measles comes much closer to satisfying Dretske’s “probability 1” criterion than that between spots in the sense of facial rash and measles. However, it is the external rash which, much like with smoke or clouds, is easily detectable by humans, and thus useable as a natural sign. Dretske clearly interprets Grice as referring to an external rash:

The red spots all over Tommy’s face mean that he has the measles, not simply because he has the measles, but because people without the measles don’t have spots of that kind. (Dretske, 1988, 56)

As we will see, the final clause is simply false. Nevertheless, an external rash does indeed aid doctor’s in the diagnosis of measles, so let’s examine the relationship
between spots in this sense and measles.

Now, the external rash resulting from measles is an indicator that measles is present, but it is not a robust enough indicator to, on its own, justify the diagnosis of measles. For example, similar rashes can be caused by “Rubella; glandular fever; drug rashes—especially due to penicillin or sulphonamides—erythema multiforme” (Rendle-Short et al., 1985, 81).¹ ² More anecdotally:

The more intense forms of measles rash can lead to confusion with other erythmata. I have met with one such case in which there was marked infiltration of the individual spots, they were of a nodular form, livid red in appearance, and particularly as they stood in thick groups together, several of my colleagues made the diagnosis of variola. A glance in the mouth serves as a rule to correct the error . . .

From scarlet fever the initial symptoms of measles are distinguished by the greater affection of the alimentary tract in the former, the greater angina, and the form of the rash. The region of the lips and chin is regularly free from rash in scarlet fever. An error in regard to scarlet fever can arise with the so-called confluent measles, yet in the general grouping together of all the symptoms, and the scrutiny of all the parts affected by the rash one will soon find some point or another characteristic of measles. Measles and scarlet fever may however occur together, and then they form a difficult diagnostic puzzle.

Serum rashes must be mentioned in conjunction with that of measles as they can show a great similarity in the skin and mucous membranes. The absence of the Koplik spots, the irregularity in the outbreak of the rash, also the sequence in which the several parts of the skin are affected, and above all the fact of the injection of the serum, will overcome the difficulty as to diagnosis. I have twice seen intense large typhoid roseola spots which had a great similarity to measles. (Pfaundler and Schlossmann, 1908, 258)

¹Note: Rubella is also known as “German measles”, but it is distinct from the so-called “English measles” under discussion here. Although English measles is also known as Rubeola, the two conditions should not be confused. Most importantly, they are caused by distinct viruses and are unrelated from a medical standpoint. In fact, it is precisely the confluence of external signs (symptoms) which inspires us to call both “measles” even though the underlying causes are distinct.

²Sulphonamides are a large class of drugs based on the sulfonamide functional group which include antibiotics, anticonvulsants, and other applications. Allergic reactions to sulphonamides occur for about 3% of the population. Erythema multiforme is a rash-like skin disorder with a number of potential causes.
This passage illustrates three crucial points about the role of natural signs in medical diagnosis.

The first point is simply the importance of a confluence of signs for practical diagnosis. If one is to correctly distinguish measles from scarlet fever, one most consider “the general grouping together of all the symptoms”; certainty of diagnosis increases with the number of indicators for the disease in question.

The second point is that brute category judgments may be refined in order to improve diagnoses. If one is to distinguish the rash caused by measles from the rash caused by allergic reaction to some serum, one must consider “the irregularity in the outbreak of the rash, also the sequence in which the several parts of the skin are affected.” However, it is also important to note that such refinements are not always possible. Even coarse categories of symptoms contribute essential information for diagnosis.

The third point is the effect of additional knowledge on the informational value of a natural sign. If one knows that a patient exhibiting a measles-like rash has recently been injected with penicillin, one should seriously consider the possibility that the rash is the result of an allergic reaction rather than the measles virus.

So, on the one hand background knowledge and category refinement contribute to our use of natural signs. On the other hand, even though the correlation doesn’t hold with probability 1, the appearance of a rash greatly aids in the diagnosis of measles.
Bibliography


